The Fordham Intellectual Property,
Media, and Entertainment Law Journal
27th Annual Symposium

Platform Society: Copyright, Free Speech, and
Sharing on Social Media Platforms

Friday
October 4, 2019

8:30 – 9 a.m. | check-in
9 a.m. - 3:30 p.m. | program

Fordham Law School | Costantino Room

CLE COURSE MATERIALS

This program is presented in conjunction with
Table of Contents

1. Speaker Biographies

   McIlwain, Charlton. *Racial formation, inequality and the political economy of web traffic.* (View in document)

   Barocas, Solon; Selbst, Andrew D. *The Intuitive Appeal of Explainable Machines.* (View in document)

2. CLE Materials

   **Panel 1: Copyright Regimes & Automated Content Moderation**

   University of South Hampton. *Better Regulation for Copyright.* (View in document)

   Mazziotti, Giuseppe Dr.. *Remuneration of Content Creators in the Digital Space.* (View in document)

   Gingras, Richard. *Proposed copyright rules: bad for small publishes, European consumers and online services.* (View in document)

   Engstrom, Evan; Feamster, Nick. *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools.* (View in document)

   Vincent, James. *This is what Google says search will look like under EU copyright Laws.* (View in document)

   **Panel 2: Human Dignity, Agency, & Algorithmic Bias**

   University of Pennsylvania Law Review. *Accountable Algorithms.* (View in document)

   Barocas, Solon; Selbst, Andrew D. *Big Data’s Disparate Impact.* (View in document)

   Barocas, Solon. *Losing Out on Employment Because of Big Data Mining.* (View in document)

   Angwin, Julia; Parris Jr., Terry.. *Facebook Lets Advertisers Exclude Users by Race.* (View in document)

   Robertson, Adi, The Dextriarchy. *Facebook ad delivery could be inherently discriminatory, researchers say.* (View in document)

   Robertson, Adi, The Dextriarchy. *What happens next in the housing discrimination case against Facebook?*. (View in document)

   Kaminski, Margot E.; Selbst, Andrew D.. *The Legislation That Targets the Racist Impacts of Tech.* (View in document)

   **Panel 3: Platform Governance**

   Douek, Evelyn. *Facebook’s “Oversight Board:” Move Fast With Stable Infrastructure and Humility.* (View in document)

   Manhattan Community Access Corp. Et Al. v. Halleck et al. (View in document)


   Gorwa, Robert. *What is Platform Governance?* (View in document)

   Douek, Evelyn. *Why were members of congress asking Mark Zuckerberg about Myanmar? A Primer.* (View in document)

   Douek, Evelyn. *YouTube’s Bad Week and the Limitations of Laboratories of Online Governance.* (View in document)

   **Key Note Session**

   Supporting Documents Only- No CLE credits offered

   Klonick, Kate; Kadri, Thomas. *How to Make Facebook’s ‘Supreme Court’ Work.* (View in document)

Brian Carver  
Copyright Counsel, Google LLC  

Brian Carver is Copyright Counsel at Google where he advises on copyright matters across the company. He was formerly Product Counsel at YouTube where he also focused on copyright matters including the notice and takedown process and copyright management tools such as Content ID. Previously, he was on the faculty at the University of California, Berkeley where he taught intellectual property and cyberlaw courses. He has also litigated intellectual property cases as an associate at a Silicon Valley-based law firm. Brian received his J.D. from UC Berkeley Law and is a member of the State Bar of California. He is originally from Birmingham, Alabama.

Nora Choueiri  
Senior Legal Counsel, Dailymotion  

Nora Choueiri is Senior Legal Counsel at Dailymotion in New York where she handles all variety of legal matters for the company, including advising on copyright matters. She was formerly the Dean’s Fellow for the Fashion Law Institute at Fordham University School of Law. Nora received her B.A. in Political Sciences with distinction from Cornell University, and her J.D. cum laude from Fordham University School of Law. While at Fordham Law she was a member of the Samuelson-Glushko Intellectual Property and Information Law Clinic and was an author in an amicus brief to the second circuit for the case Viacom Intern., Inc. v. YouTube, Inc., 676 F.3d 19 (2012). Nora is a member of the State Bar of New York. She was born and raised in New York City.

Rebecca Crootof  
Ph.D., Assistant Professor of Law, University of Richmond School of Law  

Rebecca Crootof is an Assistant Professor of Law. Dr. Crootof’s primary areas of research include technology law, international law, and torts; her written work explores questions stemming from the iterative relationship between law and technology, often in light of social changes sparked by increasingly autonomous systems, artificial intelligence, cyberspace, robotics, and the Internet of Things. She is interested in the various ways both domestic and international legal regimes respond to and shape technological development, particularly in the armed conflict context. Dr. Crootof earned a B.A. cum laude in English with a minor in Mathematics at Pomona College; a J.D. at Yale Law School; and a PhD at Yale Law School. She consults for the Institute for Defense Analyses and is a member of the New York Bar, the Equal Rights Center’s Board of Directors, the Center for New American Security’s Task Force on Artificial Intelligence and National Security, and the Permanent Mission of the Principality of Liechtenstein to the United Nations’ Council of Advisers on the Application of the Rome Statute to Cyberwarfare. She enjoys skiing, hiking, and reading science fiction “for work.”

Robert J. deBrauwere  
Partner and Co-Chair, Digital Media Group, Pryor Cashman LLP  

Partner Robert deBrauwere co-chairs Pryor Cashman’s Digital Media Group and is a member of the Intellectual Property, Media + Entertainment and Litigation Groups. His extensive knowledge of the interplay between IP, technology, media, entertainment and music has earned him the role of trusted business and legal advisor to some of the world’s most cutting-edge companies and influential artists and songwriters.

A former computer programmer and concert and theater sound/lighting engineer, Rob understands how technology and entertainment intersect in the digital world. He recognizes the challenges that accompany the design and development of new applications, products and distribution channels, and works alongside his clients to devise solutions that are scalable and secure.

Evelyn Douek  
S.J.D. Candidate, Harvard Law School  

Evelyn Douek is an S.J.D. candidate at Harvard Law School and Affiliate at the Berkman Klein Center For Internet & Society. She studies international and transnational regulation of online speech and content moderation institutional design. Her supervisors are Professors Martha Minow, Jack L. Goldsmith and Cass R. Sunstein. Prior to coming to HLS, Evelyn was an Associate (clerk) to the Honourable Chief Justice Susan Kiefel of the High Court of Australia. She also worked in commercial litigation in Sydney. She graduated with First Class Honours from UNSW Sydney, where she was the Executive Editor of the
UNSW Law Journal and the Undergraduate Student Representative on UNSW Council, the University's governing body. Evelyn blogs regularly at Lawfare, and her research has appeared or is forthcoming in the North Carolina Journal of Law and Technology, the Federal Law Review, Slate, Mother Jones, and a number of other publications.

**Abner S. Greene**  
Leonard F. Manning Professor of Law, Fordham University School of Law (Moderator)


**Diane L. Houk**  
Of Counsel, Emery Celli Brinckerhoff & Abady LLP

Diane L. Houk joined Emery Celli Brinckerhoff & Abady LLP as Counsel in 2009. Her practice focuses on representing individual and organizational plaintiffs in housing discrimination matters, including Broadway Triangle Community Coalition v. New York City; National Fair Housing Alliance v. Facebook, Inc; Fair Housing Justice Center v. M&T Bank; Kneer and Long Island Housing Services v. German American Settlement League, Inc., and L. C. v. LeFrak City. Prior to joining the firm, Diane co-founded the New York City-based Fair Housing Justice Center in 2004 and served as its first Executive Director until 2009.

Ms. Houk previously worked at the Civil Rights Division of the U.S. Department of Justice in the Housing & Civil Enforcement Section from 1991 to 2004. During her tenure as a Senior Trial Attorney, Ms. Houk served as lead counsel in more than thirty housing discrimination cases filed by the United States in federal courts throughout the country. Her docket included systemic “pattern and practice” cases, as well as individual complaints referred by HUD, involving allegations of race, national origin, familial status, sex, and disability discrimination. These included rental discrimination cases based on “testing” investigations and local land use cases, such as United States v. Yonkers, NY; United States v. Parma, OH; United States v. City of Jacksonville, FL; and United States v. Pooler, GA.

In 2000, Ms. Houk was named Special Litigation Counsel by the Division. In that role, she oversaw the development and litigation of race, national origin, and religious discrimination cases involving land use and zoning, redevelopment plans, building and occupancy codes, and affordable housing programs.

**Kate Klonick**  
Ph.D., Assistant Professor of Law, St. John’s University School of Law

Kate Klonick is an Assistant Professor at St. John's University Law School and an Affiliate Fellow at Yale Law School’s Information Society Project. Her research on networked technologies' effect on social norm enforcement, freedom of expression, and private governance has appeared in the Harvard Law Review, New York Times, New Yorker, The Atlantic, The Guardian and numerous other publications.

**Ron Lazebnik**  
Clinical Associate Professor of Law, Fordham University School of Law (Moderator)

Ron Lazebnik is the Director of the Samuelson-Glushko Intellectual Property and Information Law Clinic as well as the J.D. Externship program at Fordham University. He has assisted clinic students with a number of significant legal matters, including the filing of amicus briefs in important IP cases before the U.S. Court of Appeals for the Second Circuit. In 2012, under his leadership, the clinic became a part of the U.S. Patent and Trademark Office pilot programs to allow students to practice before patent and trademark examiners. His academic and scholarly interests include IP law, information law, and Internet law.

**Giuseppe Mazziotti**  
Ph.D., Assistant Professor of Law, Trinity College Dublin

Giuseppe Mazziotti is an Assistant Professor in intellectual property law at Trinity College Dublin as of October 2014. He held the same position at the University of Copenhagen from 2009 to 2011. Giuseppe was Visiting Scholar at the University of California, Berkeley (2004/2005), Columbia Law School (2010/2011), Pompeu Fabra University and ESADE Business School in
Jennifer Pariser is the Vice President, Copyright and Legal Affairs at the Motion Picture Association of America (MPAA). Prior to joining the MPAA, Jenny previously served as head of litigation at the RIAA and at Sony Music where she oversaw all litigation matters including significant lawsuits against online copyright pirates such as Napster, Grokster, Limewire as well as individual infringers. Prior to that she was an associate with the firms Patterson, Belknap, Webb & Tyler and Debevoise & Plimpton in New York and also served as a judicial clerk to the Honorable Charles Tenney in the United States District Court for the Southern District of New York.

Jennifer Pariser is the Vice President, Copyright and Legal Affairs at the Motion Picture Association of America (MPAA). Prior to joining the MPAA, Jenny previously served as head of litigation at the RIAA and at Sony Music where she oversaw all litigation matters including significant lawsuits against online copyright pirates such as Napster, Grokster, Limewire as well as individual infringers. Prior to that she was an associate with the firms Patterson, Belknap, Webb & Tyler and Debevoise & Plimpton in New York and also served as a judicial clerk to the Honorable Charles Tenney in the United States District Court for the Southern District of New York. She graduated from New York University Law School where she was a member of the Law Review and the School of Foreign Service at Georgetown University.
She lectures extensively on copyright topics including at the Copyright Society of the USA, the American and New York Bar Associations, PLI and various law schools.

**Pamela Samuelson**
Richard M. Sherman Distinguished Professor of Law, Berkeley Law School, Co-Director Berkeley Center for Law & Technology; Bacon-Kilkenny Visiting Professor of Law, Fordham University School of Law

Pamela Samuelson is the Richard M. Sherman ’74 Distinguished Professor of Law at the University of California at Berkeley and a Co-Director of the Berkeley Center for Law & Technology. For the Fall Term 2019, she is the Bacon-Kilkenny Visiting Professor of Law at Fordham Law School. She teaches courses on intellectual property, cyberlaw, and information privacy. She has written and spoken extensively about the challenges that new information technologies pose for traditional legal regimes, especially for intellectual property law. She is a member of the American Academy of Arts & Sciences, a Fellow of the Association for Computing Machinery (ACM), a Contributing Editor of Communications of the ACM, a past Fellow of the John D. & Catherine T. MacArthur Foundation, and an Honorary Professor of the University of Amsterdam. She is a co-founder and chair of the Board of Directors for Authors Alliance, a nonprofit whose mission is to facilitate authorship in the public interest. She is also Chair of the Board of Directors of the Electronic Frontier Foundation, as well as a Fellow of the Center for Democracy & Technology, and a member of the Advisory Boards for Public Knowledge and the Electronic Privacy Information Center.

A 1971 graduate of the University of Hawaii and a 1976 graduate of Yale Law School, Samuelson was a litigation associate with the New York law firm Willkie Farr & Gallagher before turning to academic pursuits. From 1981 through June 1996 she was a member of the faculty at the University of Pittsburgh Law School. She has been a member of the Berkeley faculty since 1996 as well as a Visiting Professor at Columbia, Cornell, Emory, Harvard, and NYU Law Schools.

**Andrew Selbst**
Postdoctoral Scholar, Data & Society Research Institute (Moderator)

Andrew Selbst is a Postdoctoral Scholar at Data & Society and Visiting Fellow at Yale Law School’s Information Society Project. Selbst studies the effects of technological change on legal institutions and structures, with a particular focus on how technology disrupts society’s traditional understandings of civil rights and civil liberties. His current research examines how certain standard legal concepts that serve as underlying bases for accountability, such as explanations, fault, and liability, may need to be reexamined as applied to machine learning systems.

Before joining Data & Society, Selbst was a Visiting Researcher at Georgetown University Law Center and a scholar in residence at the Electronic Privacy Information Center. Prior to that, he has been a senior associate at Hogan Lovells, a Supreme Court Assistance Project Fellow at Public Citizen, and a Privacy Research Fellow at NYU’s Information Law Institute. He clerked for Hon. Jane R. Roth of the U.S. Court of Appeals for Third Circuit, and Hon. Dolly M. Gee of the U.S. District Court of the Central District of California. Selbst earned his J.D. at the University of Michigan Law School and holds M.Eng. and S.B. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. He is licensed to practice law in New York, New Jersey, and Washington, D.C.

**Ana-Andreea Stoica**
Ph.D. Candidate, Computer Science Department, Columbia University

Ana-Andreea Stoica’s work focuses on mathematical models, data analysis, and policy implications for algorithm design in social networks. From recommendation algorithms to the way information spreads in networks, Ana is particularly interested in studying the effect of algorithms on people’s sense of privacy, of community, and of access to information and opportunities. She strives to integrate tools from mathematical models - from graph theory to opinion dynamics - with sociology, to gain a deeper understanding of the ethics and implications of using technology in our everyday lives. Ana grew up in Bucharest, Romania, and moved to the U.S. for college, where she graduated from Princeton in 2016 with a bachelor’s degree in Mathematics.

**Nadine Strossen**
John Marshall Harlan II Professor of Law, New York Law School

Nadine Strossen, a chaired professor at New York Law School, is a leading expert and frequent public speaker/media commentator on constitutional law and civil liberties issues. The immediate past President of the American Civil Liberties Union (1991-2008), she serves on the national advisory boards of the ACLU, Electronic Privacy Information Center, Foundation for Individual Rights in Education (FIRE), and Heterodox Academy.
The National Law Journal has named Strossen one of America’s “100 Most Influential Lawyers,” and several other national publications have named her one of the country’s most influential women.

BETTER REGULATION FOR COPYRIGHT

ACADEMICS MEET POLICY MAKERS

KEYNOTE by Commissioner Mariya Gabriel

PANEL DEBATES: • Neighbouring right for publishers • Platform Liability • Copyright on data

WED 6 SEPT 2017 15:00–18:30

European Parliament Room ASP1G3

Please RSVP: juliareda.eu/better
Contents

Neighbouring Right for Publishers

The proposed press publishers' right: an actual solution? 5
The proposed publishers' right in press publications: an evidential mistake 11
The Press Publishers' Right in a Nutshell 17

Platform Liability

Filtering obligations and fundamental rights: can the EU eat the cake and have it too? 23
Addressing the value gap on user-generated content platforms from the perspective of weaker copyright holders 31
Form and Substance in the Value Gap Proposal 39

Copyright on Data

Data producer’s right: Powers, Perils and Pitfalls 51
Text & Data mining Issues 59
Data Property: Unwelcome Guest in the House of IP 65

All contents are under Creative Commons CC-BY licence, unless otherwise specified
Neighbouring Right for Publishers
The proposed press publishers’ right: an actual solution?
Eleonora Rosati*

Abstract

Article 11 of the draft Directive on copyright in the Digital Single Market contains a provision that, if adopted in the form proposed by the European Commission, would introduce a new neighbouring right at the EU level in favour of press publishers for the digital use of their press publications.

The proposal has attracted significant commentary. This brief note discusses whether – from a copyright perspective – the idea of an EU-wide press publishers’ right: is supported by an internal market rationale which justifies an intervention at the EU level; will grant press publishers broader and more certain protection than the one already enjoyed under the EU copyright acquis; will improve press publishers’ ‘bargaining position’ as per the Commission’s stated intention.

Overall, the answer appears to be in the negative. This contribution holds the view that – at best – a press publishers’ right will not change the situation of its beneficiaries and – at worst – will increase the complexity of the legal system and distract the attention from other options that could be potentially more effective in supporting the European press publishing sector.

The content of the proposal

Under the umbrella of its Digital Single Market Strategy1 and among a number of other legislative proposals, in the final part of 2016 the European Commission released a proposal for a new directive on copyright in the Digital Single Market2 (‘DSM Directive’).

With the declared goal of helping press publishers “increase their legal certainty, strengthen their bargaining position and have a positive impact on their ability to license content and enforce the rights on their press publications”3, the draft DSM Directive contains a provision which, if adopted in the form proposed by the Commission, would introduce a new neighbour-ing right over press publications for their digital use. The rationale of the proposal stems from awareness of the difficulties facing press publishers when seeking to license their publications and prevent unauthorized uses by online services.4


---

*Associate Professor in Intellectual Property Law, University of Southampton. Email: eleonora@e-lawnora.com.


rightholders, in respect of the works and other subject-matter incorporated in a press publication. The duration of this new neighbouring right would expire 20 years after the publication of the press publication.6

**Critics of the proposal**

The Commission’s proposal for a press publisher’s right has been subject to extensive commentary and possibly even more extensive criticism, notably within academic circles. Having already expressed critical views regarding the introduction of a press publishers’ right at the EU level further to some national experiences – notably Germany (sections 87f, 87g and 87h of the Urhrechterrechtsgesetz, ie the German Copyright Act)) and Spain (Art. 32 of the Ley de Propiedad Intelectual, ie the Spanish Intellectual Property Law)7 – at the time (2016) of the Commission’s Consultation on the role of publishers in the copyright value chain8, in early 2017 the European Copyright Society (ECS) referred once again to the proposal for a neighbouring right in favour of press publishers in negative terms. Overall sceptical regarding the actual achievement of its underlying goal (this being to support a struggling newspaper industry), the ECS also stated that an exclusive right to control the exploitation of press contents online would “not only negatively affect freedom of expression and information, but also distort competition in the flourishing European information market.”9 The latter would be because of higher barriers of entry to the online news market that would make it more difficult for emerging businesses to access it.

The position of the ECS echoes similar views expressed in late 2016 by a group of 37 intellectual property professors based in the UK, and the Opinion the Centre for International Intellectual Property Studies (CEIPI) at the University of Strasbourg.

In a letter sent to the Copyright Policy Directorate of the UK Intellectual Property Office, a group of 37 professors from a number of UK universities considered that the Commission’s proposal for a press publishers’ right would be “unnecessary, undesirable, would introduce an unacceptable level of uncertainty and be unlikely to achieve anything apart from adding to the complexity and cost of operating in the copyright environment.”10

6 The full text of Article 11 is as follows:

**Article 11 Protection of press publications concerning digital uses**

1. Member States shall provide publishers of press publications with the rights provided for in Article 2 and Article 3(2) of Directive 2001/29/EC for the digital use of their press publications.

2. The rights referred to in paragraph 1 shall leave intact and shall in no way affect any rights provided for in Union law concerning the use of the works and other subject-matter incorporated in a press publication. Such rights may not be invoked against those authors and other rightholders and, in particular, may not deprive them of their right to exploit their works and other subject-matter independently from the press publication in which they are incorporated.


4. The rights referred to in paragraph 1 shall expire 20 years after the publication of the press publication. This term shall be calculated from the first day of January of the year following the date of publication.

The Commissioners for a press publishers’ right would fail to contribute to the construction of a Digital Single Market; be contrary to the interests of authors; be supported by a clear economic rationale; be detrimental to the public domain; do not contribute to any publication, including those in respect of which the relevant press publisher’s investment has not been substantial; and have an excessive duration.12

**The Parliamentary debate**

Further to the release of the Commission’s proposal for a DSM Directive, the discussion moved to the European Parliament, where MEP Therese Comodini Cachia was appointed rapporteur on behalf of the Committee on Legal Affairs (JURI).

On 10 March 2017 a first draft of her report on the proposed DSM Directive was released.13 MEP Comodini Cachia appeared to take a rather radically different view regarding the desirability of having an EU-wide press publishers’ right. In fact, she proposed that press publishers would be granted, not a neighbouring right over their press publications, but rather (and more simply) a presumption of representation of authors for the sake of rights enforcement (Amendment 52):

“Member States shall provide publishers of press publications with a presumption of representation of authors of literary works contained in those publications and the legal capacity to sue in their own name when defending the rights of such authors for the digital use of their press publications.”

In mid-2017 MEP Comodini Cachia announced that she was renouncing her role at the European Parliament, and MEP Axel Voss was appointed new rapporteur on the proposed DSM Directive. The final version of the Report is scheduled for adoption in the final part of 2017.

In the meantime, other Parliamentary committees have expressed views for the Committee on Legal Affairs on the proposed DSM Directive. In its Opinion (Rapporteur: MEP Catherine Stihler) the Committee on the Internal Market and Consumer Protection (IMCO) held the view that there are no sufficient grounds that would justify the introduction of a press publishers’ right.14 To ameliorate enforcement of rights, it would be sufficient to amend Article 5 of Directive 2004/48/EC15 (the ‘Enforcement Directive’) by means of a regulation that would make this provision also applicable to press publishers. The Opinion also holds the view that “[t]here are potentially more effective ways of promoting high-quality journalism and publishing via tax incentives instead of adding an additional layer of copyright legislation.”16

Other Committees have also proposed amendments to the original Commission’s proposal, especially for the sake of clarifying the scope of the resulting right. In its draft Opinion (Rapporteur: MEP Marc Joulaud), the Committee on Culture and Education (CULT) deemed it necessary to clarify that non-commercial and private uses of professional press publications are not covered17, and that protection does not extend to acts of hyperlinking, or to the text fixating the hyperlink, where such acts do not constitute communication to the public under the InfoSoc Directive.18

With a similar tone, the CEIPI Opinion considers that the proposal for a press publishers’ right would likely fail to contribute to the construction of a Digital Single Market; be contrary to the interests of authors; be supported by a clear economic rationale; be detrimental to the public domain; likely fail to contribute to any publication, including those in respect of which the relevant press publisher’s investment has not been substantial; and have an excessive duration.19

17 Committee on Culture and Education, Draft opinion on Legal Affairs on the proposed DSM Directive. In its Opinion (Rapporteur: MEP Catherine Stihler) the Committee on the Internal Market and Consumer Protection (IMCO) held the view that there are no sufficient grounds that would justify the introduction of a press publishers’ right.14 To ameliorate enforcement of rights, it would be sufficient to amend Article 5 of Directive 2004/48/EC15 (the ‘Enforcement Directive’) by means of a regulation that would make this provision also applicable to press publishers. The Opinion also holds the view that “[t]here are potentially more effective ways of promoting high-quality journalism and publishing via tax incentives instead of adding an additional layer of copyright legislation.”16
18 Other Committees have also proposed amendments to the original Commission’s proposal, especially for the sake of clarifying the scope of the resulting right. In its draft Opinion (Rapporteur: MEP Marc Joulaud), the Committee on Culture and Education (CULT) deemed it necessary to clarify that non-commercial and private uses of professional press publications are not covered17, and that protection does not extend to acts of hyperlinking, or to the text fixating the hyperlink, where such acts do not constitute communication to the public under the InfoSoc Directive.18
In a similar fashion, the Opinion (Rapporteur: MEP Zdzisław Krasnodębski) of the Committee on Industry, Research and Energy (ITRE) recommended the inclusion of a new recital that would state that “[t]he rights for press publishers should apply without prejudice to the rights of individuals for the reproduction, communication or providing links or extracts of a press publication to the public for private use or not-for-profit, non-commercial purposes.” 20

Three questions

While awaiting further developments at the level of EU legislation, this note will only consider whether – from a copyright perspective – the idea of an EU-wide press publishers’ right:

• is supported by an internal market rationale which would require an intervention at the EU level;

• will grant press publishers broader and more certain protection than the one already enjoyed under the existing acquis;

• will improve press publishers’ "bargaining position".

Overall, the answer appears to be in the negative. While this brief note does not touch upon potential issues connected with fundamental rights (notably freedom of expression and information) and competition law, it holds the view that the adoption of a press publishers’ right is unlikely to change the situation of press publishers. It could however increase the complexity of the legal system and distract the attention from other options that could be potentially more effective in supporting the European press publishing sector.

(1) An internal market rationale?

Similarly to the other EU copyright directives, the legislative basis for the proposed DSM directive is Article 114 of the Treaty on the Functioning of the European Union (TFEU), ie the realisation of an internal market where the free circulation of goods and services based on or incorporating copyright content is ensured. All this is premised upon the idea that differences in Member States’ laws are such as to raise barriers to such free circulation. As in all cases of shared competence, such as copyright and more generally intellectual property, the directive must also satisfy the requirements of subsidiarity (Article 2(2) TFEU) and proportionality (Article 5 of the Treaty on European Union).

With regard to the proposed press publishers’ right, neither the draft directive nor the accompanying Impact Assessment (‘IA’) provide a satisfactory explanation as to why intervention at the EU level is needed. In particular, the IA provides contradictory inputs. It recalls that a number of Member States has already intervened to remedy or reduce – whether by means of ad hoc initiatives or as part of broader arrangements – the impact of a failure to establish a harmonised system of copyright exceptions and limitations across the EU. 21

Focusing specifically on the German and Spanish experiences, ie recent initiatives specifically mentioned at Recital 34 of the draft DSM Directive for quotation and news reporting, it holds the view that the adoption of a (non-waivable) fair compensation requirement, in Spain.

The new neighbouring right will be also subject to copyright exceptions and limitations as is the case in Spain. 21

20 Committee on Industry, Research and Energy, Opinion <TitreType></TitreType><CommissionResp>on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market, 2016/0280(COD) , Article 18.</TitreType>


22 Ibid, 161.


24 See further E Rosati, ‘Copyright in the EU: in search of (in)flexibilities’ (2014) 9(7) JIPLP 585, 590-594, questioning to what extent diverging implementations of Article 5 exceptions and limitations are allowed by the InfoSoc Directive.


26 See further, Rosati, ‘Neighbouring rights’, cit. 569-570 and 573-574.
lishing industry and included in the Commission’s Impact Assessment, it cannot be inferred that the positive trends associated with growth of digital revenue are due to the existence of a copyright environment in which press publishers benefit from an ad hoc right.27

Conclusion

From the brief analysis conducted above it would appear that the Commission’s proposal on a press publishers’ right is not firmly supported by an internal market rationale. It is also unlikely to improve the position of press publishers substantially, possibly with the exception of enforcement scenarios in which a less pervasive measure, eg a presumption or representation, would suffice.

Lacking a clear basis that justifies the introduction of a new neighbouring right at the EU level from a copyright perspective, other types of solutions – also indicated in the various parliamentary committees’ opinions – could be explored to support the press publishing sector, whether at the national or EU levels.

The proposed publishers’ right in press publications: an evidential mistake

Dr. S.J. van Gompel, Institute for Information Law (IViR), University of Amsterdam

Introduction

One of the most controversial features of the European Commission’s proposal for a Directive on copyright in the Digital Single Market is the provision introducing a related (or ‘ancillary’) right for publishers of press publications (art. 11 CDSM proposal).2 As it is currently proposed, this provision would grant publishers of press publications a set of broad exclusive rights of reproduction and communication to the public to authorise digital uses of their press publications until 20 years after first publication, subject to the same exceptions and limitations that apply to copyright works. Effectively, it would mean that, unless an exception or limitation applies, prior authorization would have to be obtained from publishers for any digital reproduction (direct or indirect, temporary or permanent, by any means and in any form) and any making available of their press publications, in whole or in part, including possibly the smallest snippets.3 This right is offered in addition to existing copyrights protecting the content (articles, photographs, illustrations, etc.) of newspapers, magazines, journals and other periodicals.

In the past year, fierce criticism has been raised against the proposed publishers’ right, both by academics,4 independent publishers5 and other stakeholders, including creators in the news publishing industry.6 In this contribution, the key points of criticism will be analysed and discussed in the light of the EU’s objective for ‘better regulation’. After a short introduction into the background of the proposal, the paper will elaborate on four main objections against the proposed publishers’ right. It will conclude that, in view of the evidence available, it is clear that the proposal is ill-suited to address the problems that press publishers are facing. Therefore, the proposed publishers’ right should at best be removed from the legislative agenda or at worst be replaced by a presumption that publishers represent the authors’ copyright in press publications and have the right to sue in their own name against digital infringement of that copyright, as was proposed in the draft report of the European Parliament’s JURI committee of 10 March 2017.

1. The research for this paper was conducted in the framework of the research programme Veni with project number 451-14-033 (‘The challenge of evidence-based intellectual property law reform: Legal pragmatism meets doctrinal legal reasoning’), which is partly financed by the Netherlands Organisation for Scientific Research (NWO).


3. See the Impact Assessment accompanying the proposal (SWD(2016) 301 final, Brussels, 14 September 2016), p. 157 (n. 485), referring specifically to the CJEU’s Infopaq I judgment (Case C-5/08), in which it was held that capturing 11-word text fragments of newspaper articles constitutes a reproduction in part of these works under art. 2(a) Directive 2001/29 ‘if the elements thus reproduced are the expression of the intellectual creation of their author’.


**Background to the proposal**

A central point that the Commission wishes to address by the introduction of the press publishers’ right is the future sustainability of the quality press, which according to the Impact Assessment is in jeopardy. As this ‘would be prejudicial for the media pluralism, good quality information and the role [press publishers] play in democratic societies’, the Commission believes that legislative intervention at EU level is needed.

In a nutshell, the problems that press publishers are facing stem from the fact that they have been struggling to cater to the two-sided market of readers and advertisers in the digital environment. In recent years, press publishers have seen a significant decline in print readership due to structural changes in consumer behaviours. In the past, it were traditional outlets such as newspapers, radio and TV channels that brought news to the people, but nowadays, most news is consumed on the internet, through different digital formats and online sources. Data provided by the press publishing sector show a steady decline in print circulation of daily newspapers in eight EU Member States, although the differences between countries are noticeable, varying from an 8% decline in Belgium, to an 18% decline in the UK and a 52% decline in Italy in the period 2010-2014.

Concomitantly, press publishers have seen structural changes in advertising markets. Advertising takes place where audiences can best be reached. As a consequence, online advertising has grown at the cost of traditional off-line advertising. This has affected news publishers in particular, as advertisers tend to focus on search engines, social media and other channels over news media. News publishers have also lost their position in the advertising market for jobs, housing, (used) cars, classifieds, which on the internet is controlled predominantly by specialised platforms and online marketplaces.

As a result of these developments, news publishers have witnessed a persistent decline in turnover over the past years, both in terms of sales and advertising revenues, which is expected to continue in the near future. This has already caused news publishers to close down or reduce editorial staff, thus leading to a decline of quality of the free and pluralist press. If, due to their poor financial situation, press publishers can make less resources available to conduct quality journalism, they may indeed lose ‘gatekeeping’ power. This threatens the traditional function of the press as a ‘public watchdog’ and may put citizens’ access to information at risk. Ultimately, such state of affairs could be detrimental to public debate and the proper functioning of a democratic society.

To ensure the sustainability of a free and pluralist quality press, news publishers have called for a new ‘ancillary right’ that enables them to (a) take legal action against online infringements of their publications, and (b) license their publications to online service providers, such as social media, news aggregators and search engines, which currently provide unauthorized access to press publications made freely available online by news publishers. This has resulted in the proposed press publishers’ right, which is aimed at protecting the investments of publishers in producing press publications.

**Key objections against the proposal**

**Objective 1: A publishers’ right is unnecessary as press publications are already protected**

It is somewhat awkward that the Commission is proposing a new related right in press publications, the content of which normally already benefits from copyright protection. Most press articles, photographs, illustrations, etc. are protected by copyright, which is usually transferred to press publishers before publication. Accordingly, press publishers often enjoy copyright protection in their press publications due to a transfer of rights by journalists, photographers, illustrators, etc. Press publishers nevertheless complain that licensing and enforcement in the digital environment is complex and inefficient, as they are not recognised as rightholders in their own right. But this raises the question: Why would existing copyright not be a good enough instrument to protect the interests of press publishers? And why would they be helped by the introduction of an additional layer of rights, which essentially grants a similar type of protection?

The Commission maintains that the introduction of a self-standing intellectual property right in press publications is needed to tackle the legal uncertainty that press publishers face when licensing and enforcing rights in the online environment. But that argument cannot convince. Although it may be easier for press publishers to negotiate licenses if they have their own right, they can already license on the basis of the copyright that is contractually obtained from journalists and other content creators. To the extent that press publishers face difficulties to prove that they own the copyright in press articles (i.e. to establish the chain of title of all rights in their publications), the legal uncertainty they face is unmistakably the result of a lack of adequate rights administration and does not warrant the introduction of a new press publishers’ right.

**Objective 2: The proposed right does not fix the problems of the press**

Although sometimes met with scepticism, the problems that news publishers are facing with the transition from print to digital are real and should be taken seriously. They might warrant legislative action, but the idea that introducing a press publishers’ right would help to cure the existing problems of print media in the digital environment is mistaken. Clearly, neither the behaviour of news consumers nor the advertising market will change as a result of the introduction of a press publishers’ right. Accordingly, the proposal by no means addresses the key underlying drivers of the problem.

Moreover, while the Commission assumes that the proposed press publishers’ right will have a positive effect on media pluralism, the relationship between the two is unclear. In general, it is difficult to establish a causal effect between intellectual property rights and incentives to invest in content creation, let alone to demonstrate that a publishers’ right will aid media pluralism. Even if it would yield additional income for publishers, it cannot be automatically assumed that the money will be invested in journalistic efforts. Hence, there is no evidence that the introduction of a press publishers’ right will result in better news coverage or the creation of more diverse media content.

The proposal may even have adverse effects on media pluralism, as it is uncertain how online service providers will respond to the introduction of a press publishers’ right. If they will refuse to engage in licensing negotiations with publishers and stop providing access to newspaper contents, as Google News and other news aggregators initially did in Germany and Spain where

---

18 M.H. van Eechoud, ‘A publisher’s intellectual property right: Implications for freedom of expression, authors and open content policies’, study conducted on commission from OpenForum Europe, January 2017, par. 2.3.
19 Impact Assessment, op. cit., Annex 13A.
21 Impact Assessment, op. cit., Annex 13A.
22 Id., p. 156 (reporting on outcomes of the 2016 public consultation on the role of publishers in the copyright value chain).
similar, though narrower, rights in press publications have been introduced, this may have negative effects on the accessibility of news online and will certainly lead to a fall in referral traffic to newspaper websites. This may be harmful for small press publishers, in particular. A 2017 study shows that, after the introduction of the obligation to pay compensation for online use of news articles in Spain, the traffic to Spanish newspaper websites fell by a 5.3% decline in visits on average, with a decline of 4.9% for large newspapers, 6.3% for medium-sized newspapers, and 12.6% for small newspapers. As a result, press publishers attract significantly less advertising revenue, which in Spain is estimated to be around 9-18 million annually in the short term. The EU legislator should not take such effects lightly, but examine them seriously before even considering to introduce a press publishers’ right. This is particularly important in light of new business models in online news publishing, which are still in development, as the Commission also acknowledges. Caution is warranted, as it is uncertain how the introduction of a press publishers’ right will affect traditional as well as future business models, including the B2B licensing market for online news publications.

Objection 3: The proposed right is possibly bad for authors of press publications

An additional concern is that the proposed press publishers’ right might have a negative impact on journalists, photographers, illustrators and other creators, whose works are included in news articles. Although the proposal clearly states that the press publishers’ right ‘shall in no way affect any rights provided for in Union law to authors and other rightholders, in respect of the works and other subject-matter incorporated in a press publication’, it cannot be excluded that it will nevertheless affect them. This is especially the case for journalists, photographers, illustrators and other creators who work as freelancers. To establish a name and reputation, which is crucial for their work and business, freelancers need maximum exposure of their work online. A press publishers’ right might hinder that. As Van Eechoud explains: ‘If the operation of the proposed publisher’s right were to lead to a decline in referrals, shares, snippet-linking or the ability to blog about a journalist’s works, this would directly harm the journalist’s visibility, and thus opportunity to sell future work.’

Also, the proposal may worsen the bargaining position of journalists and other content creators. There is no guarantee that, after the press publishers’ right is introduced, more money will become available to compensate for the online use of press articles. If the pie would grow, the surplus will presumably be taken by press publishers in the exercise of their related right. If the pie remains the same, there is a reasonable chance that press publishers on the basis of their related right will demand a larger share of it, in which case journalists, photographers and other creators would need to take a loss.

Objection 4: The proposal is overly broad

Apart from the questionable assumptions that underlie the proposal, there are further ambiguities, in particular regarding its beneficiaries and scope. Even if the proposal would aid press publishers, it goes way further than what is required to protect the quality press. The definition of ‘press publication’ (art. 2(4) CDSM proposal) is so broad that it covers virtually all content published periodically under a heading of news. The beneficiaries of the proposed right are thus not only the quality press, but all periodical media, including newsletters, blogs, glossies, social media, etc. It goes without saying that not all these media play an equally important role in democratic societies and they certainly are not in a similar situation of crisis as the quality press. In respect of its scope, it is unclear why the proposal affects all online users of news, including consumers and other legitimate users, if the real intention is to target the use of press articles by social media, news aggregators and search engines. Furthermore, do press publishers really need broad exclusive rights, if they merely seek to participate in the advertising revenues generated by the content on their websites? And why is a 20-years term of protection proposed, if the commercial life span of most press articles is no longer than a day, a week or a month at most? These are all questions to which the Commission’s proposal does not provide adequate answers.

The proposal further leaves unanswered what exactly will be protected: would the right protect press publications, including the content they comprise, or merely the fixation of press publications as identifiable media items? In the latter case the right seems useless, as online service providers seldom reuse or provide access to media items as a whole, but rather offer snippets to their content. If the right would also protect the content of a press publication, however, then this might extend the scope of protection beyond that of copyright protection. Brief and simple news items that contain little expression apart from facts, such as ‘news of the day’ or ‘miscellaneous facts having the character of mere items of press information’ are outside the scope of copyright, but would arguably be protected under the proposed press publishers’ right. Not only would this be contrary to the Berne Convention, but it would also impair the free flow of information.

Conclusion

The proposal for a related right for publishers of press publications is flawed. Although the problems that print media are facing in the online environment are real, there is no evidence that a press publishers’ right will meaningfully contribute to addressing these problems. Also, there is genuinely no need for a new right in press publications, as news publishers often already benefit from copyright protection contractually obtained from journalists and other content creators. In response to the draft impact assessment that accompanies the proposal, the European Commission’s Regulatory Scrutiny Board also observed that ‘[t]he report should more convincingly demonstrate that the creation of a new standalone right for news publishers would effectively contribute to reinforcing their role in the digital world and that action at EU level is needed.’ In my view, it still makes no case for why the introduction of a press publishers’ right is needed.

Various European Parliament’s committees that are looking into the matter also seem to recognize that the evidence is against the proposal. The CULT Committee’s draft report of 6 February 2017 advised to significantly limit the proposal and the draft report of the IMCO Committee of 20 February 2017 even suggested to abandon the press publishers’ right altogether. The suggestion to replace the right with a presumption that publishers represent the authors’ copyright in press publications and have the right to sue in their own name against digital infringement.

22 S&H-F. German Copyright Act grants press publishers a one-year exclusive right to prohibit the use of ‘snippets’ (not being individual words or the smallest of text excerpts) by search engines and news content aggregators; Art. 32(2) Spanish Intellectual Property Act contains an obligation for content aggregators to pay compensation to news publishers for the use of ‘snippets’ of press publications.


24 Id., p. 62.


27 Van Eechoud 2017, op. cit., p. 5.
of that copyright, as was made in the draft report of the JURI Committee of 10 March 2017, is probably easier to reach political agreement on. But in the absence of final votes on their positions, all is still out in the open.

Accordingly, the European Parliament has an important task ahead to make right what is wrong. Admittedly, this task is not easy. Still, the EU legislator should be very cautious to create a new right without having a clear picture of all its intended and unintended consequences. Lawmaking is not a process of trial-and-error. It actually has bearing on the subjects targeted by the legislation. As the potential impacts of the proposal on the position of journalists, on media pluralism, on future business models and on the B2B licensing market are not yet assessed, thoughtfulness and caution are warranted.

---

**The Press Publishers’ Right in a Nutshell**

Thomas Höppner*

1. Current Market Failure Necessitating the Proposed Right

The proposed right reacts to new technical opportunities for the mass-copying of press publications combined with strong economic incentives for companies to take advantage of these technical opportunities.


Some 20 years ago, when the current InfoSoc directive 2001/29 was drafted, there was no need for an independent publishers’ right. It was simply not economically viable for any company to copy and distribute newspapers and magazines en masse. Today, as a result of digitalisation, this is a different story of course. Press publications can be replicated and distributed globally through various digital platforms in the blink of an eye.

2. Strong Economic Incentives to Copy Press Publications en masse

There are not just new means to mass-copy, there are also strong economic incentives for companies to do so. This incentive is inherent in the internet ecosystem. It constitutes the standard business model of the internet economy to publish attractive content on one’s website in order to attract internet users for advertising or subscription purposes. The easiest way to do that, of course, is to take the content from other websites and to display it on one’s own site. The economic success of such aggregation platforms depends on bundling and presenting as much content as possible. More content attracts more users and more users mean higher advertising revenues or subscription fees. Consequently, every aggregator has an economic incentive to display as much third-party content as possible directly on its site.

This also applies to news intermediaries. There are many examples for such aggregators. They systematically index and copy third party news websites in order to set up their own news outlet. To this end, they have typically pre-installed news categories on their homepages. Users are presented with the most relevant text extracts that are often sufficient to convey the key message and thus the value of the article. If a user clicks on any of the news extracts copied from third party news websites, even more third party content becomes visible. Users are then invited to comment or to otherwise interact with the aggregating website. Thus, aggregators use third party press publications to build up own monetisable customer relationships and to ultimately keep users away from the source and on their own website. That explains why according to the Commission’s Impact Assessment, 47% of all users of aggregators do not click through to a press publishers’ site anymore.1

3. Aggregators’ Free-riding on Press Publishers’ Efforts Distorts Competition and Eliminates Incentives to Invest and Innovate

Now, if nearly half of the users of such aggregators do not click through to publishers’ sites but remain on the aggregator’s site, it is apparent to me that we are not talking about a symbiosis here between aggregators and press publishers or even a win-win-situation as has been claimed by others. Instead, many aggregators are directly competing with press publishers’ sites for the same advertising budgets by satisfying the same information demand of the same users. Regrettably, at the moment, they are also competing with the same content that the press publishers produce at high costs and that the aggregators merely copy at no costs. This is a classical market failure that needs to be addressed. The legal framework has to ensure that press publishers’ incentives to invest in reliable, high-quality publications are maintained. That in turn is a traditional function of copyright law. The publisher’s right as proposed by the EU Commission is capable of addressing this issue.

---

* Prof Dr Thomas Höppner, professor of business and IP law at Technical University Wildau, visiting professor of law at University of Strathclyde, Glasgow; partner with law firm Hausfeld LLP, Berlin.

II. The Proposed Right Strikes the Right Balance of Interests

Contrary to all the scaremongering from the side of aggregators, there are no substantial reasons for anyone to be afraid of the proposed right. It strikes a fair balance.

1. The proposed right does not limit but strengthens consumers’ access to news

First, it has been argued that the publishers’ right somewhat limits the flow of information and the ability of consumers to express their opinions freely. However, it is difficult to see how the proposed right could have such effects. Press publishers vigorously defend the freedom of expression. Why would they want to prevent their readers from engaging with their publications? That is neither the purpose nor any likely effect of the publishers’ right. The right does not protect any ideas or facts that are published. Everyone remains free access to the press publications published online. And everyone remains free to comment on any subject covered by a publication, to quote it, to link to it or to share it with others. Recital 33 of the proposed directive clarifies that the proposed right “does not extend to acts of hyperlinking” which are typically used for sharing and referencing purposes.

2. The proposed right supports a continuing availability of quality news

The critics ignore that the alternatives to a publishers’ right would interfere much more significantly with the freedom of expression and consumers’ access to information than a publishers’ right. Without a better protection of publishers’ online investments, press publishers would be forced to either invest less in quality content, make less content available online or to hide that content behind paywalls and subscription models. Each of these alternatives would leave consumers worse off. It is the consumers who benefit the most from the press publishers’ current approach of making news available for free and anyone online. Aggregators’ unrestricted copying of this content is threatening this approach.

3. The proposed right does not harm but back up journalists by empowering the entire press

Another point of criticism is that the publishers’ right could in some way harm journalists by decreasing their public exposure. Again, it is difficult to see how this could happen. It would be surprising if there was any journalist who would not prefer being paid to merely being visible online but unpaid. The publishers’ right aims at securing the sustainability of the entire press including its journalists. Empowering the press with an updated legislation is the best measure politicians can do to secure a diverse and open media landscape.

4. The narrow definition of a “press publication” appropriately restricts the rights’ scope

Another criticism that has been raised is that the right was apparently too wide as it had no built-in restrictions. In fact, the proposed right contains a very important built-in restriction, that is the definition of the term “press publication”. To be protected, the publication must be, inter alia, “a fixation of a collection of literary works of a journalistic nature within a periodical or regularly-updated publication under the initiative, editorial responsibility and control of a service provider.” These criteria are not easy to fulfill. They require substantial and continuous investment and justify particular protection.

5. The proposed right is narrower than that for other media publishers

The proposed publishers’ right is not wider but in fact narrower than comparable related rights for music producers, film producers or broadcasting organisations. Their protection is granted for certain activities, namely for the mere first technical fixation of a phonogram, a film or a broadcast, irrespective of the quality, relevance or originality of these activities. There is no legitimate reason to treat press publications any differently. If anything, considering their relevance for democratic societies, press publications merit additional legal protection not less. That is why the proposed publisher’s right merits support.

---

3 Van Eechoud, A publisher’s intellectual property right: Implications for freedom of expression, authors and open content policies, OpenForum Europe 2017.
4 ibid.
Platform Liability
Filtering obligations and fundamental rights: can the EU eat the cake and have it too?
Sophie Stalla-Bourdillon*

Setting the problem

The European Commission in its explanatory memorandum to the proposed new Copyright Directive in the Digital Single Market1 released on 14 September 2016 states that the proposal “has a limited impact on the freedom to conduct a business and on the freedom of expression and information, as recognised respectively by Articles 16 and 11 of the Charter, due to the mitigation measures put in place and a balanced approach to the obligations set on the relevant stakeholders.”2 The European Commission, however, does not explain how the mitigation measures and its balanced approach meet fundamental rights requirements and does not even address the risk that its proposal could lead to divergent interpretations across and within Member States. Notably, the European Commission does not even refer to all the fundamental rights taken into account by the Court of Justice of the European Union (CJEU) when undertaking its own balancing exercise, including the rights to data protection and privacy (which were considered in the Promusicae/Sabam3 and Scarlet/Sabam4 cases).3 The European Commission’s impact assessment is not of great help either in this respect.4 The European Commission seems to rely on the fact that, implicitly, the choice of restriction is left, in the first instance, to the service providers meaning that it is not directly imposed by the transposing legislation.

It is therefore crucial to determine whether the proposed Copyright Directive strikes an appropriate balance between the different fundamental rights at stake in the light of CJEU case law and in particular, whether Article 13 is compatible with the EU acquis broadly defined including the Charter of Fundamental Rights of the European Union (the EU Charter).

Article 13 of the proposed Copyright Directive in its first paragraph provides that:

“Information society service providers that store and provide to the public access to large amounts of works or other subject-matter uploaded by their users shall, in cooperation with rightholders, take measures to ensure the functioning of agreements concluded with rightholders for the use of their works or other subject-matter or to prevent the availability on their services of works or other subject-matter identified by rightholders through the cooperation with the service providers. Those measures, such as the use of effective content recognition technologies, shall be appropriate and proportionate. The service providers shall provide rightholders with adequate information on the functioning and the deployment of the measures, as well as, when relevant, adequate reporting on the recognition and use of the works and other subject-matter.”

It thus expressly prescribes the use of effective content recognition technologies as a means to ensure the functioning of agreements with rightholders or to prevent the availability of copyright works. Recital 38 is more explicit in that it specifies that the obligation to “take appropriate and proportionate measures to ensure protection of works or other subject-matter, such as implementing effective technologies” should also apply when the information society service providers are eligible for the liability exemption provided in Article 14 of Directive 2000/31/EC.”

* Associate Professor in Information Technology Law, University of Southampton, UK.
3 Case C276/06 Promusicae, Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), 29 January 2008, EU:C:2008:54, para. 64.
4 Case C-70/10, Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), 24 November 2011, EU:C:2011:771, e.g. para. 50.
5 For a similar point see e.g. Stalla-Bourdillon S et al. (2016), A brief exegesis of the proposed Copyright Directive, https://ssrn.com/abstract=2875296.
In other words, hosting providers, as a species of intermediary providers, should also implement effective technology, by which one should understand content recognition technology, also implementable upon upload filters.

Value creation results more generally by transformative uses of copyright works.13 The chain of value creation to which online platforms contribute does not end with online platforms. By way of example, are diverted from known applications to be deployed in new contexts.7 "Value creation results from actions that entail the novel combination and exchange of resources, by which resources are appropriated part of the value generated by the use of copyright works. Several meanings can be attached to the concept of "value. "Economic theory suggests that the value of copyright works is not and should not be entirely appropriated by rightholders. The market value of a copyright work and its societal value are measured differently and should evolve over time. However, value creation also relates to innovation. Value creation results in the novel combination and exchange of resources, by which resources are diverted from known applications to be deployed in new contexts."7 The European Commission therefore does not account for the value created by online platforms themselves and more generally by transformative uses of copyright works.12 The chain of value creation to which online platforms contribute does not end with online platforms. By way of example, plenty EU companies develop their businesses based on code-sharing platforms.14 Furthermore, value created by online platforms does not necessarily relate to the use of copyright works but also to the use of personal data relating to service users (in particular data about content consumption and social sharing habits) for service personification and advertising purposes in the context of highly concentrated markets. As a result, Article 13 as a means of redistributing value among platforms and rightholders can only be a very approximate tool unable to cater for complex value creation processes with multiple stakeholders.

Upload filters are a reality. Google’s Content ID18 is probably the most well-known example. However, this is an expensive technology and only few online platforms use them.16 YouTube claimed in 2016 to have invested more than $60 million in Content ID.17 Could the technology be then realistically licensed for a few hundred euros? Ironically, making upload filters mandatory would thus strengthen the market position of dominant online platforms. This is also supported by the fact that neither Article 13 nor Recitals 38 or 39 clarify whether online platforms will be permitted to use the reference files corresponding to copyright right without requesting a license.

Even if upload filters are already a reality, the EU legislature should be asking whether upload filters should be made mandatory, for a significant number of market players and for all types of protected works (e.g. including code-sharing platforms and software).

A crucial reason why making upload filters a requirement for certain online platforms is an ill-conceived policy is its incompatibility with fundamental rights. This has been explained by the CJEU in a series of cases starting with the landmark Scarlet/Sabam case and followed by the Sabam/Netlog case.18 In Netlog, in particular, the CJEU clearly highlighted the intimate relationship that exists between a cornerstone of the digital single market legal framework, i.e. Article 15 of the E-commerce Directive,19 and the protection of fundamental rights.

In Sabam/Sabam, the Belgian Association of Authors, Composers and Publishers, had requested "that Netlog be ordered immediately to cease unlawfully making available musical or audio-visual works from SABAM’s repertoire and to pay a penalty of EUR 1000 for each day of delay in complying with that order."20 How could Netlog, a social media platform, comply with the order in practice? By using content recognition technology and implementing upload filters. Netlog then claimed "that the granting of such an injunction could result in the imposition of an order that it introduce, for all its customers, in abstracto and as a preventative measure, at its own cost and for an unlimited period, a system for filtering most of the information which is stored on its servers in order to identify on its servers electronic files containing musical, cine-

Assessing mandatory upload filters in the light of fundamental rights

Mandatory upload filters are the wrong answer to a poorly formulated problem, confusingly referred to as the “value gap” conundrum.8 The European Commission in its memorandum explaining the new proposed copyright Directive states that “[i]t is therefore necessary to guarantee that authors and rightholders receive a fair share of the value that is generated by the use of their subject-matter.16 Implicit in this statement is the idea that online platforms contribute to the public access to large amounts of works or other subject-matter uploaded by their users.24 Could it be that such a requirement would prove incompatible with the EU acquis and the protection of fundamental rights? To say it bluntly, are mandatory upload filters and fundamental rights friends or foes? Can the EU legislature impose ante filtering obligations upon certain online platforms and still argue that the proposed Copyright Directive is fundamental rights compatible? More prosaically, can the EU legislature eat the cake of fundamental rights and have it too?

### 7 As per Article 13 of the proposed Copyright Directive.


### 9 Proposed Copyright Directive, p. 3.

### 10 C. Handke, Y. Girard, A. Mattes (2015), Fördert das Urheberrecht Innovation? Eine empirische Untersuchung, Stu-

### 11 Case 256/11 Sabam v Netlog, Final Report - The Economic and Social Impact of Software and Services on Competitiveness and Innovation, (SMART 2015/0015), p. 12. Notably, many independent projects of a small or medium size rely upon code-sharing platforms. For example, the Open Source Initiative Content ID work, https://opensource.org/doc/content-id/index.html, has "Videos uploaded to YouTube are scanned against a database of files that have been submitted to us by content owners. Copyright owners get to decide what happens when content in a video on YouTube matches a work they own. When that happens, the video gets a Content ID claim. Copyright owners can choose to block or monetise the video or track its viewership statistics." 19 One could also add to the list Audible Magic technology, https://www.audiblemagic.com/Contri

### 12 The market value of a copyright work and its societal value are measured differently and should evolve over time. However, value creation also relates to innovation. Value creation results from actions that entail the novel combination and exchange of resources, by which resources are diverted from known applications to be deployed in new contexts.7 The European Commission therefore does not account for the value created by online platforms themselves and more generally by transformative uses of copyright works.12 The chain of value creation to which online platforms contribute does not end with online platforms. By way of example, plenty EU companies develop their businesses based on code-sharing platforms.14 Furthermore, value created by online platforms does not necessarily relate to the use of copyright works but also to the use of personal data relating to service users (in particular data about content consumption and social sharing habits) for service personification and advertising purposes in the context of highly concentrated markets. As a result, Article 13 as a means of redistributing value among platforms and rightholders can only be a very approximate tool unable to cater for complex value creation processes with multiple stakeholders.

Upload filters are a reality. Google’s Content ID18 is probably the most well-known example. However, this is an expensive technology and only few online platforms use them.16 YouTube claimed in 2016 to have invested more than $60 million in Content ID.17 Could the technology be then realistically licensed for a few hundred euros? Ironically, making upload filters mandatory would thus strengthen the market position of dominant online platforms. This is also supported by the fact that neither Article 13 nor Recitals 38 or 39 clarify whether online platforms will be permitted to use the reference files corresponding to copyright right without requesting a license.

Even if upload filters are already a reality, the EU legislature should be asking whether upload filters should be made mandatory, for a significant number of market players and for all types of protected works (e.g. including code-sharing platforms and software).

A crucial reason why making upload filters a requirement for certain online platforms is an ill-conceived policy is its incompatibility with fundamental rights. This has been explained by the CJEU in a series of cases starting with the landmark Scarlet/Sabam case and followed by the Sabam/Netlog case.18 In Netlog, in particular, the CJEU clearly highlighted the intimate relationship that exists between a cornerstone of the digital single market legal framework, i.e. Article 15 of the E-commerce Directive,19 and the protection of fundamental rights.

In Sabam/Sabam, the Belgian Association of Authors, Composers and Publishers, had requested "that Netlog be ordered immediately to cease unlawfully making available musical or audio-visual works from SABAM’s repertoire and to pay a penalty of EUR 1000 for each day of delay in complying with that order."20 How could Netlog, a social media platform, comply with the order in practice? By using content recognition technology and implementing upload filters. Netlog then claimed "that the granting of such an injunction could result in the imposition of an order that it introduce, for all its customers, in abstracto and as a preventative measure, at its own cost and for an unlimited period, a system for filtering most of the information which is stored on its servers in order to identify on its servers electronic files containing musical, cinc-
A careful analysis of these three arguments shows that it is the unfettered monitoring of all or most of the information stored by the online platform that raises fundamental rights concerns: either because it amounts to the systematic processing of personal data, or because it is unable to distinguish between lawful and unlawful information or more simply because of its costs. There is in particular a very intimate relationship between the finding of a monitoring obligation of a general nature and the finding of an engagement of Article 8. Certain amendments have been tabled in the European Parliament arguing that only the reference files are being checked, and thereby there is no processing of personal data. However, personal data must be processed in order to allow the proposed appeals process to function. It is also argued that a search for specific reference files is not a general monitoring. This ignores the fact that it is the size of the targeted population that is crucial (i.e. whether the entire user base is in fact targeted by the monitoring or not) as well as the wide range of information searched for. Furthermore, the number of reference files will run into hundreds of thousands, if not millions. Assuming each search for a matching with a reference file should be deemed monitoring “in a specific case,” within the meaning of Recital 47 of the E-commerce Directive, which in itself is debatable as all service users would be concerned as well as all their platform activities, Article 15 would certainly be undermined by the systematic juxtaposition of hundreds of thousands of specific cases. Finally, the argument based on cost should be related to the discussion on innovation and value creation processes.26

In consequence, the E-commerce Directive and its Article 15 are not the only reason why Article 13 of the proposed copyright directive is problematic. Said otherwise, fundamental rights concerns explain why mandatory upload filters are incompatible with Article 15. Article 15 of the E-commerce Directive thus finds its roots in the protection of fundamental rights, including the rights to data protection and freedom of expression of Internet users.

Reforming copyright at the EU level

What should the EU legislation do then? Once again, the decision of the CJUE Sabam/Netlog is worth reiterating and in particular the reasons for finding that Article 11 of the EU Charter is engaged in the case at hand. The CJEU observes, more or less implicitly, that beyond the approximation of upload filters the route to automation can only reasonably start with the harmonisation of copyright exceptions. To repeat the words of the CJEU, “it is not contested that the reply to the question whether a transmission is lawful also depends on the application of statutory exceptions to copyright which vary from one Member State to another.” Going back to the big sister of the proposed Copyright Directive, i.e. the infosoc Directive,27 the latter comprises 20 optional exceptions and thereby cannot be describe as an attempt to reduce divergences.28 The three mandatory exceptions29 introduced in the proposed Copyright Directive will not change the situation.

For freedom of expression concerns to be (partially) alleviated, context assessment should be made possible in order to determine if the act at stake is permitted or justified without the rightholder’s consent. Because the technology itself is not able to assess context, processes in the sense of a set of steps taken in order to reach a decision whether to restrict access to a particular content become crucial to meet fundamental rights requirements.

---

26 See p. 3.
29 See Articles 3-5 of the proposed Copyright Directive.
Article 13 does not put in place a proper process for the use of content recognition technology in order to reach a decision whether to restrict access to particular content. The only safeguard found in Article 13 would become relevant once the decision to restrict access has been taken, i.e. the availability of complaints and redress mechanisms.30 Notice-and-action procedures are one example of processes, although there is a range of variations of notice-and-action procedures that are well balanced. However, the EU legislature has up until now always postponed the task to harmonise through hard law or soft law such procedures, although such a possibility had been envisaged right from the beginning by the drafters of the E-commerce Directive.31

Notice-and-action procedures are all the more important since the right to freedom of expression is not the only fundamental right at stake. The right to data protection in the EU Charter and the right to the respect of one's private life imply following CJEU case law that the systematic processing of all or most user information to prevent copyright infringement raise serious concerns. Yet notice-and-action procedures are based on the premise that the processing undertaken by rightholders is only partial as per definition rightholders do not have access to the entirety of user information. In addition, notice-and-action procedures make possible the coupling of a second stage consisting in an assessment of the context in which the copyright work at stake is actually being borrowed from. This does not mean that notice-and-action procedures cannot be partially automated,32 at least at the detection stage. As a result, the use of content recognition technology should not be equated to the implementation of upload filters.

Amending Article 13, Recitals 38 & 39

In its draft report,33 rapporteur Therese Comodini Cachia proposes to amend Article 13 and Recitals 38 and 39. The amended version of Article 13 does not refer to content recognition technology anymore. Because of the confusion widely spread between the use of content recognition technology and the implementation of upload filters this disappearance could be seen as an improvement of the text. Besides, it is now clarified that intermediary providers would not be required to implement upload filters, although it is not entirely clear who would count as “information society service providers [that] are actively and directly involved in the making available of user uploaded content to the public and where this activity is not of a mere technical, automatic and passive nature.”34 Nevertheless, the reference to “appropriate and proportionate measures to ensure the functioning of agreements concluded with rightholders for the use of their works” is maintained. Whether such a drafting is enough to make it clear to both national legislatures and judges that mandatory upload filters should not become a requirement could still be questioned.

With this said, the new paragraph 2a adds that:

“The measures referred to in paragraph 1 shall be implemented without prejudice to the use of works made within an exception or limitation to copyright. To this end, Member States shall ensure that users are allowed to communicate rapidly and in an effective manner with the rightholders who have requested the measures referred to in paragraph 1 in order to challenge the application of those measures.”

30 Article 13(2) of the proposed Copyright Directive provides that: “Member States shall ensure that the service providers referred to in paragraph 1 put in place complaints and redress mechanisms that are available to users in case of disputes over the application of the measures referred to in paragraph 1.”
31 See Article 2(2) of the E-commerce Directive.
Addressing the value gap on user-generated content platforms from the perspective of weaker copyright holders

Dr Giuseppe Mazziotti - Trinity College Dublin

Introduction

In a piece published on The Guardian in October 2013, the former leader of Talking Heads, David Byrne, was very pessimistic about how the Internet would have impacted on the commercial value of copyright. Byrne argued that streaming services such as Spotify and Pandora might be good to discover new music but they bring no income to today’s artists across the creative industries. Byrne’s voice has not been isolated in emphasizing a situation that is due not only to the notoriously weak bargaining power of the average authors and performers vis-à-vis content producers (e.g., record labels, film studios, etc) but also to the widely uncompensated dissemination of copyright works on user generated content platforms and social media.1

From the outset of the so-called ‘Web 2.0’, the preservation of the incentive/reward rationale of copyright on online platforms that encourage and enable user creativity and participation has been a highly debated and complex problem. It is still unclear how copyright works and materials that users share and make available to others without the authorization of the respective rights-holders should be identified and removed. More than ten years after the first publication of a YouTube video,2 this question is still valid and open in Europe because of the legal uncertainties that characterize the regime of liability of providers of such services in different jurisdictions.

A copyright reform proposed by the European Commission in September 2016 seeks to bring the aforementioned uncertainties to an end.3 Notwithstanding its title (i.e. ‘Copyright in the Digital Single Market’) the priority of the draft directive is not the attainment of territorial unification or integration of digital markets. At least when it comes to the business-related aspects of the draft directive, the main goal is, rather, the achievement of a well-functioning market place for creative works through a EU-wide, more effective and broader protection of copyright. This objective is pursued through several provisions that aim at bringing greater fairness and transparency in the assignment and enforcement of copyright, also in the domain of user generated content. As acknowledged in the impact assessment that accompanies the directive proposal, licensing practices in this sector have been very difficult to develop. One of the main assumptions the directive proposal has drawn upon is that the advent of user-generated platforms and social media in the last decade has triggered a significant erosion of the value of copyright because of the central role these platforms have gained in making creative works available to the public and given the widely unlicensed character of the copyright works they provide access to.4

1 David Byrne, ‘The Internet will suck all creative content out of the world’, The Guardian, 11th of October 2013.
2 Byrne mentioned the example of the song of the summer 2013, ‘Get lucky’ by Daft Punk, which made the two members of the band earn approximately 18,000 USD each, as a result of 104,760,000 Spotify streams this track reached until the end of August 2013 (“What happens to the bands who don’t have International summer hits?”, he commented).
3 IFPI & IMPALA (coordinators), Securing a sustainable future for the European music sector, Letter addressed to Jean-Claude Juncker on 29th of June 2016 (signed by almost 1300 artists and songwriters from across Europe as of 18th of July 2016). In the letter the artists claimed that the future of music was jeopardized by a substantial ‘value gap’ caused by user upload services like Google’s YouTube, which are taking value away from the music community and from its artists and songwriters.
4 The first video was uploaded on YouTube on the 23rd of April 2005 by Jawed Karim (with the title “Me at the zoo”).
6 The central role of such companies and of their services is also due to their very successful businesses: YouTube’s value was estimated to be $70 billion and its revenues in 2015 were reported to be $9 billion; Pinterest has been valued $12 billion in 2015; Soundcloud $700 million in 2014, etc. These figures are reported in European Commission, Impact assessment on the modernisation of EU copyright rules, Commission Staff Working Document, SWD(2016)301 final, Part 1/3, Brussels, 14.09.2016, p. 146 (hereinafter ‘Impact assessment’).
From a legal point of view, the debate has been widely monopolised by the question about whether and how these kinds of platforms should be granted the special immunity from liability Directive 2000/31 (e-Commerce Directive) created for providers of hosting services when the web was in its infancy.7 The objective of Article 13 of the current draft Directive is seeking to bring clarity on this front, entail that interactive services which play an active role in giving access to user generated content (e.g., by optimising the presentation of the uploaded works or promoting them) shall be subject to copyright and be obliged to take measures to make unauthorised contents inaccessible.8

As we will see, this means that platforms such as YouTube, Facebook, Twitter, Instagram, Pinterest, Vimeo and Soundcloud and many others would be expected to clear copyright for the contents that pop up on their platforms and to implement technologies which ensure an accurate identification of those items made available by their service subscribers without the permission of copyright holders.

**The ‘value gap’ provision**

Article 13 of the draft directive aims at creating an obligation for providers of user-generated online content platforms to prevent the availability of unauthorised works on their services by means of appropriate and proportionate measures, for instance by implementing effective technologies. In the copyright reform debate and in the policy documents used for the preparation of the legislative proposal the provision has been constantly associated to a ‘value gap’ (from an expression used for the first time by the music sector) that would exist to the detriment of the creative sector because of the widely uncompensated use of copyright works.9

The entities that would be obliged to make un-licensed works inaccessible are “… providers that store and provide to the public access to large amounts of works or other subject matters” as clarified by the Commission, assessing the size (i.e. “large amounts”) of a user generated content platform would require an analysis of combined factors, such as the number of users and visitors and the amount of content uploaded over a given period of time.10 Appropriate and proportionate measures to be implemented would include technologies - such as Google’s Content ID, currently used across the YouTube platform - ensuring accurate identification of the works for which the platform devisers have concluded licensing agreements. However, to be fully implemented it has been claimed that the text of the Commission proposal could be better drafted in so far as it does not specify whether Article 13 would impose a new filtering obligation only on platforms with existing licensing agreements or - as it seems more correct, at least to me - on all platforms, regardless of these agreements.11

In essence, the provision at issue aims at re-affirming the basic copyright principle according to which whoever reproduces and/or communicates to the public works protected by copyright must obtain the permission to do so from their respective copyright holders. Without such permission - under copyright law - unauthorised uses such as those happening on social media or online platforms certainly expose users to the consequences of copyright infringement. However, it still uncertain in the case law of the EU member states whether or not the platform deviser would be (directly or indirectly) liable for infringements of their subscribers.12 A limited liability of platforms has somehow been entailed by judgments of the Court of Justice of the European Union (CJEU). In particular, the Court found that the unauthorised embedding of copyright works on social media is lawful in so far as the linked or embedded content has already been made available to the public with the consent of the copyright holder.13 The CJEU found also that the installation of permanent filtering measures on social networks for purposes of copyright enforcement would be excessively costly and burdensome for online platform and would eventually run against their freedom to do business online.14

From a formal point of view, Article 13 would not directly modify the regime of liability exemption granted to providers of ‘hosting’ services. The e-Commerce Directive, at the end, would not be amended. The draft text, in this regard, clarifies that Article 13 would be applicable also to providers who qualify for the existing exemption, in a way that the two layers of regulation would operate simultaneously and in a complementary manner. However, it is evident that the provision would avoid, with EU-wide effects, an excessively broad application of the ‘safe harbour’ exemption by not allowing national courts to shield online platforms from liability in case of online copyright infringement. What is controversial is the fact that the provision would change the allocation of responsibilities between copyright holders and service providers, with the new law giving the exact view of how copyright technologies - such as Google’s Content ID - would have to be treated them differently from how we treat professional content providers, such as media companies? If such services, in addition to user creations, systematically give us access to contents not remunerating what their businesses rely upon? Are the platforms media companies “by accident”? If the new provision were interpreted, as it seems fair, as presupposing the obligation for online platforms to enter into licensing agreements with copyright holders, the implementation of such a regime would have generated a new ‘value gap’ (from an expression used for the first time by the music sector) that would exist to the detriment of the creative sector because of the widely uncompensated use of copyright works. From a formal point of view, Article 13 would not directly modify the regime of liability exemp-

8 See the 2016 directive proposal, recital 38.
9 Helen Smith, (IMPALA), Véronique Desbrosses (GESAC), Frances Moore (IFPI), ‘Value gap is crucial for the music user profiling businesses.
11 See the Open Letter of European Research Centres, p. 6.
13 All these cases are mentioned in the Commission’s impact assessment, p. 143.
17 See the Open Letter of European Research Centres, p. 6.
The European Commission’s reasoning

In the impact assessment that accompanies the directive proposal, the European Commission pointed out that a legislative intervention on user-generated content and social networks is needed for copyright holders to re-gain control over the modes and conditions of online exploitation of their works. The Commission emphasized the necessity to continue to support and remunerate digital content creation at a time when, with the rise of interactive online environments, copyright holders have proven to be unable to rely on an undisputed legal obligation for platform and service providers to obtain permission for what their users/subscribers make available online and what the platforms gain through customized commercializations.18

A situation of uncertainty, as regards such obligation, has arisen with regard to providers of social media and social network services because of their frequent inclusion, in the case law developed in various European jurisdictions, into the categories of hosting providers who benefit from the liability exemption under Article 14 e-Commerce Directive. The Commission acknowledged that this situation of uncertainty has significantly weakened the bargaining power of copyright holders and decreased the value for copyright content, especially with regard to TV broadcasts.19 On the one hand, online platform providers have offered to rights-holders – according to the impact assessment – mostly ‘monetization agreements’, which are regarded as being concluded on a voluntary basis, not as a result of an obligation of online platforms to clear copyright. On the other hand, these deals do not reflect the value (and the price) of the licensing agreements that copyright holders enter into with on-demand content suppliers such as Spotify, Deezer and Netflix and with TV broadcasters, who are therefore placed in a disadvantaged position vis-à-vis the user generated content platforms.20

To motivate its decision to intervene, the Commission stressed also that the activities of providers of today’s user generated content are unlikely to fall within the scope of the liability exemption created under Art. 14 of the e-Commerce Directive. The main reason for that is that the exemption, as explained by the Court of Justice of the European Union in a landmark decision,21 shall be reserved to technical and automatic activities, and not to services whose functionalities include content categorization, recommendations, playlists or the ability to share contents.22 In short, these functionalities make online platforms clearly distinct from the passive and unaware hosting providers the drafters of the e-Commerce Directive had in mind when shaping the exemption.

Aren’t users already obliged to not publish unauthorized works?

The rise to prominence of video-sharing platforms such as YouTube, Daily Motion and Vimeo, social networks like Facebook and Twitter and other interactive services or dedicated platforms for photos (e.g. Instagram, Pinterest) and sound recordings (e.g. Soundcloud) has significantly expanded the opportunities for Internet users to access copyright works and to become authors themselves. As of October 2015, YouTube had 1.3 billion users (i.e. one third of all Internet users),23 who see content every hour; Daily Motion has 300 million users watching 3.5 billion views every month; Vimeo has a monthly audience of approximately 170 million users and 35 million registered users; Soundcloud’s user community has grown exponentially, going from 11 million users in 2011 to 150 million in 2015 and 250 million in 2016.24 From a legal point of view, access to such platforms is made conditional upon the acceptance by each single user/subscriber of terms and conditions that are relevant from a copyright-related perspective. Standard terms and conditions that apply to online content platforms contractually oblige subscribers not to share and publish contents created by third parties and which they cannot lawfully dispose of.25 This means that no ‘value gap’ provision would have been necessary if the online platform devisers had effectively enforced their own terms of service with regard the upload of unauthorized materials. No copyright infringement would have massively materialised in the era of Web 2.0 if the online platforms, from the outset, had paid attention (and technological requirements) to the contents entire user community was sharing on their networks. In this regard, what the value gap provision is seeking to achieve now is to make legally explicit what has already been expressed and required contractually, but not properly enforced, to the detriment of the value of unauthorised creative works.

A lose-lose situation for copyright holders

It is easy to understand why user-generated content platforms and social networks give rise to a ‘lose-lose’ situation for copyright holders if one considers the two aspects below:

1. Copyright holders have not been able to rely on a proper enforcement of their rights when someone else makes available their works without permission within an online platform. If an author or a copyright holder has no resources to monitor content generation on platforms and to send notices to take content down, her copyright remains ineffective unless the platform deviser acts spontaneously and removes the unauthorised work. It must be borne in mind that notice-and-takedown procedures have been mostly used by music and film majors and by their respective anti-piracy bodies, for which these industries have invested significant amounts of money across Europe. To the contrary, the notice-and-takedown mechanism has never worked properly for individual creators such as photographers, writers, composers and film or video makers and small producers because of the lack of time and money to be dedicated to online enforcement procedures.

2. The standard terms and conditions which are accepted by users of social networks and user-generated content services require subscribers to give platforms such as YouTube, Facebook and Instagram global, perpetual licences for the making available on a user-generated content platform of all contents (e.g. writings, photos, music, multi-media and audiovisual works) authored by users. This means that whoever wished to make available his or her works in order to take advantage of the remarkable exposure opportunities provided by large online platforms would instantaneously lose – especially in the absence of bargaining power, as in the case of individual creators and small-size content producers - all remuneration opportunities by the very nature of their contracts. Moreover, the “exposure” effects the online platforms have strongly advocated are primarily beneficial to performing artists (i.e. musicians, actors) who could view uncompensated viewings on user upload platforms as a way to boost their live performance businesses. However, this is not a mechanism that fits the business of non-performing authors of works such as films, videos and other multi-media works for which the making available on a user-generated content platform inevitably compromises the exploitation and remuneration opportunities for each of those works.

Addressing the value gap from the perspective of weaker copyright holders

If one considers the relevance of copyright and of its effective enforcement for the purpose to preserve and stimulate diversity of content creation, Article 13 can be easily placed in relation to other measures of the draft directive that seek to help copyright holders gain a better position to negotiate and licence their online rights and to gain adequate remuneration. From this angle, the ‘value gap’ provision can be viewed as a complementary tool to achieve a broader policy objective, which is that of ensuring a fair share of income to copyright holders who – under the current legal framework - have been unable to gain adequate revenues from online uses of their works.

For the first time in the history of European copyright policy, the 2016 draft directive targets the contractual relationships between individual creators and their assignees with the intent to introduce a common approach of transparency requirements across the EU and to strengthen the bargaining power of original copyright holders.26 To this end, Article 14 of the proposal

18 Commission’s impact assessment, p. 139.
19 Impact assessment, p. 141.
20 Ibidem.
21 See C-324/09, L’Oreal and Others v. eDay International AG and Others (2011).
22 Commission’s impact assessment, p. 138.
23 Ibidem.
24 See, for instance, YouTube’s (http://youtube.com/terms) and Facebook’s (http://facebook.com/terms) Terms of Service.
25 Commission’s impact assessment, pp. 173 ss.
26 Commission’s impact assessment, pp. 173 ss.
obliges Member States to ensure that authors and performers be given, in accordance with the specificity of each sector of the creative industries, a right to timely and adequate information on the modes of exploitation of their works and the related revenues. Moreover, Article 15 gives authors and performers the possibility of seeking contractual adjustments whenever their originally agreed remuneration proves to be disproportionately low in comparison to the earnings and benefits derived from the online exploitation of their works.

The above-mentioned provisions show that there is a clear attempt to enable copyright holders to take advantage of the expected increase of revenues produced by the implementation of Article 13. This is even more important for copyright holders with a weaker bargaining power in so far as their newly created rights to obtain sector-specific and timely information and an adequate remuneration - to be sought also through contractual adjustments - were made effectively enforceable and non-waivable on EU-wide grounds. It would be much easier for individual copyright holders to licence and price their works, also through their collecting societies, if they knew how profitable their use has been in interactive environments. This would be true not only with regard to user generated content platforms and social networks but also in the realm of on-demand services. It is therefore not surprising that, among the hundreds of amendments which have been tabled in the European Parliament, a few of them seeking to pursue a fair remuneration strategy in a more straightforward way (i.e. through the codification of a right to remuneration to be paid directly to authors and performers for the making available of their works by on-demand content providers) were eventually incorporated in the opinions sent by the Culture and Industry Committees to the Legal Affairs Committee on the 11th of July 2017.

If the final version of the directive established a clearer and stronger link between Article 13 and the rights to information and to fair remuneration of authors and performers under Articles 14 and 15, the increased output coming from user generated content could be viewed as an equal opportunity to achieve remuneration for the aforementioned amendments to seek to introduce in the domain of on-demand content deliveries. Even though the structure of the two rights would be substantively different, both proposals aim at creating an obligation to remunerate copyright holders that would be particularly important for individuals and small companies having no or little bargaining power vis-à-vis the online platforms and the on-demand content suppliers. As far as we are concerned, as we have seen, the draft directive seeks to place authors and performers in a position to better negotiate and licence their rights, also through their respective collecting societies. According to the proposed directive, contractual remedies will have to be made available each time the originally agreed remuneration of authors and performers proves to be disproportionately low in comparison to the (transparent) revenues generated by the exploitation of their works. As regards on-demand services, in addition, the amendments proposed by the Culture and Industry Committees aim at codifying a EU-wide right to remuneration to be enforced against the providers of such services and to be administered by collecting societies of authors and performers, in both the music and audiovisual sectors.

Critical remarks

The proposal to enact the provision of Article 13 has been widely criticized in the current debate in so far as it would stifle innovation and user participation in online environments and would oblige platforms to implement technologies in sectors where content identification is flawed and unreliable. Someone has also blamed the Commission for having inserted this provision in the directive having just one company and business in mind, i.e. Google’s. A brief reflection on the aforementioned critical remarks seems to be useful:

1. **Innovation and user participation:** in an environment where a high level protection of copyright is mandated under all copyright directives and regarded as an intrinsic guarantee for the creation of professional content, interpreting the duty to remunerate authors and producers of the content a business relies upon is a misconception of what intellectual property is about. It is true that the platforms that have started their activities and developed in the last decade have taken advantage of an unclear legal framework and of a loophole in widely using copyright works for free. However, this has come with a price, which is the erosion of the commercial value of creative works and the frustration of online remuneration opportunities for professional content creators. It seems obvious to me that the value of professional content should be protected as a crucial element of the ‘Digital Single Market’ strategy.

2. **Accuracy of content identification:** in certain sectors (music, for instance) content identification technologies are more developed than in others, depending also on the availability of databases copyright holders place at the disposal of online platforms to facilitate fair and proportionate filtering mechanisms. The fact that in other sectors, such as photography, such databases and technologies are not equally available is, in my view, a further reason to introduce the criterion of non-discrimination of Article 13, which would allow devisers and rights holders an incentive to co-operate and to improve filtering and content identification technologies.

3. **Re-allocation of responsibilities:** as it has been pointed out above, the value gap provision entails a different regulatory treatment for a kind of intermediation in digital content distribution which is characterised by the active role of the platform and considers the opportunities of content identification and nuanced filtering mechanisms which has become available recently. The E-Commerce Directive and notice-and-takedown procedures would still matter and remain applicable in their own sphere of application.

4. **Freedom to do business online:** the technological measures required under the value gap provision would have to be fair and proportionate and should not cause ‘over-blocking’ or restriction of free and legally unprotected materials made available by the platform users.

5. **Addresses of the provision:** it is paradoxical that the initiative of the Commission might have been taken as a measure to be enacted just against the provider of one user generated platform, i.e. YouTube, and not also as a tool to foster copyright clearance activities on multiple platforms. YouTube’s deviser, i.e. Google, was actually a pioneer in the development of content identification technologies and it is probably the platform that is best suited to achieve a well-functioning marketplace for copyright works on its platform. Due to such advancement of YouTube on content identification, it is evident that the enactment of the value gap provision would be more costly and demanding for other companies and platforms where licensing agreements and copyright enforcement measures have been poor or non-existing. The problem of the provision is actually the opposite, if one considers that the obligation to copyright works obligating the non-court use of the non-court generated platforms such as Wikipedia. In this respect, an amendment of Article 13 aimed at excluding non-for-profit platforms from the scope of application of the new provision could be a suitable solution.

**Conclusion**

After having briefly explained the rationale and policy objectives of Article 13 of the proposal for a directive on ‘Copyright in the Digital Single Market’ (the so-called ‘value gap’ provision), this paper has provided a brief reflection on the nature of the obligation to filter and make unauthorized copyright works inaccessible on user-generated content platforms. In particular, the
The 2016 draft directive embodies under Article 14 and 15 remedies (i.e. a right to information about the levels of remuneration and the modes of online accounting. The paper suggests that the provision of Article 13 could have a much stronger justification if its goal to fill a gap in the value chain of online content were more closely linked to the codification of a right to fair remuneration to be achieved through effective rights and remedies (i.e. a right to information about the levels of remuneration and the modes of online exploitation; a right to contractual redress in case of disproportionately low remuneration) that the 2016 draft directive embodies under Article 14 and 15.

Form and Substance in the Value Gap Proposal

In September 2016, the European Commission published a copyright reform package, including a new Directive on copyright in the Digital Single Market. One of the Directive’s most controversial aspects is the so-called “value gap proposal” provided in Article 13 and accompanying recitals.

This proposal suffers from a number of fundamental problems, of both a formal and substantive nature. From the formal point of view, the proposal (i) includes normative provisions in the recitals, (ii) lacks basic clarity, and (iii) uses trivially vague language. On the substantive side, if passed into law the proposal will (i) thwart digital innovation and (ii) disproportionately restrict the fundamental rights of Internet users and platform operators.

Given the seriousness of these flaws, the EU institutions should consider deleting or, at least, significantly rewriting the proposal.

Introduction

In September 2016, the European Commission published a copyright reform package. The most controversial instrument in this package is the Directive on copyright in the Digital Single Market (draft Directive). The main reason behind the controversy, along with the proposed right for press publishers, lies in the so-called “value gap proposal” provided in Article 13 and Recitals 37 to 39.

Over the last few years, rightholders and their representatives have been floating the idea that there is a value gap in the online content marketplace that is in need of closing. In their lobbying efforts, the value gap has come to replace piracy as the main digital threat to the survival of the creative industries. The idea is that rightholders face difficulties when seeking to license and be remunerated for the online distribution of their works. With the evolution of digital technologies, platforms of user-uploaded content (think of YouTube, Dailymotion and Vimeo) have become important vehicles for such distribution. According to rightholders, these platforms inappropriately invoke the “hosting safe harbour” laid down in Article 14 of the E-Commerce Directive in order to argue that they are under no obligation to conclude licensing agreements. Rightholders claim that, as a result, they are not always able to obtain a fair remuneration from platforms of user-uploaded content. The unfairness, they add, is made evident by the difference between the remuneration paid by these (typically ad-funded) platforms and that paid by subscription services, such as Spotify Premium and Deezer.

1 The title of this contribution borrows from the title of one of the most influential American law review articles of the late twentieth century: Kennedy (1976).
2 Lecturer in Copyright Law, Universidade Católica Portuguesa – Faculdade de Direito (Lisbon); LL.M., Harvard Law School (Cambridge, MA).
3 With the fundamental purpose of “modernising” EU copyright rules for the digital age, the package comprises one Regulation on copyright and related rights for online transmissions and retransmissions of television and radio programmes and one Directive on copyright in the Digital Single Market, as well as two other instruments dedicated to implementing the Marrakesh Treaty into EU law.
4 The number of amendments submitted by MEPs – almost 1.000 – may serve as an illustration of the degree of controversy. See http://en.euractiv.eu/wp-content/uploads/sites/2/2017/05/JURI-copyright-amendments.pdf.
6 Criticising the introduction of a new right for press publishers, see, e.g., Hugenholtz (2016); Kretschmer et al. (2016); Peukert (2016); Ramalho (2017); and Geiger et al. (2017).
7 The expression “transfer of value” is sometimes used instead of “value gap”. See, e.g., Lucas-Schoetter (2017).
8 See, e.g., the petition at http://www.makeinternetfair.eu.
9 See Recital 37.
Grounded on this rationale, Article 13 and Recitals 37 to 39 of the draft Directive attempt to address this gap, by reinforcing the position of rightholders to negotiate and be remunerated for the online use of their works by platforms of user-uploaded content11.

Article 13(1), the proposal’s centrepiece, reads as follows:

Information society service providers that store and provide to the public access to large amounts of works or other subject-matter uploaded by their users shall, in cooperation with rightholders, take measures to ensure the functioning of agreements concluded with right-holders for the use of their works or other subject-matter or to prevent the availability on their services of works or other subject-matter identified by rightholders through the cooperation with the service providers. Those measures, such as the use of effective content recognition technologies, shall be appropriate and proportionate. The service providers shall provide right-holders with adequate information on the functioning and the deployment of the measures, as well as, when relevant, adequate reporting on the recognition and use of the works and other subject-matter.

Since the Commission released the draft Directive, numerous independent academics have raised their voices against the value gap proposal12. With the present contribution, I merely reinforce the chorus of critics, by discussing those that are, in my view, the major problems with the proposal.

These problems are of a twofold nature: formal and substantive. Formal problems and substantive problems are naturally intertwined, in that poor drafting choices often affect the provision’s substance. For clarity of exposition, nonetheless, I will treat these problems separately. The problems, all platforms of user-uploaded content, in storing and providing public access to protected content uploaded by their users, will be prima facie liable for communicating works to the public. Abruptly, though at the same time stealthily, a significant change to the copyright acquis is proposed: if the text of Recital 36 is approved as it stands, all platforms of user-uploaded content, in providing their services, will be prima facie liable for communicating works to the public.

It is important to note that Recital 38 is not merely codifying a CJEU-developed construction13.

The departure point of the proposal seems to lie in Recital 38, which starts by stating that online platforms that store and provide public access to protected content uploaded by their users perform an act of communication to the public. Abruptly, though at the same time stealthily, a significant change to the copyright acquis is proposed: if the text of Recital 36 is approved as it stands, all platforms of user-uploaded content, in providing their services, will be prima facie liable for communicating works to the public.

1. Formal problems

1.1. The proposal includes normative provisions in the recitals

The departure point of the proposal seems to lie in Recital 38, which starts by stating that online platforms, in storing and providing public access to protected content uploaded by their users, perform an act of communication to the public. Abruptly, though at the same time stealthily, a significant change to the copyright acquis is proposed: if the text of Recital 36 is approved as it stands, all platforms of user-uploaded content, in providing their services, will be prima facie liable for communicating works to the public.

It is important to note that Recital 38 is not merely codifying a CJEU-developed construction13.

On the one hand, throughout a long line of case law, the CJEU has shaped a complex concept of communication to the public, composed of many interdependent criteria. None of these criteria, nor the way in which they interact with each other, are reflected in the value gap proposal. Instead, Recital 38 simply states that platforms go beyond the mere provision of physical

facilities and violate the right of communication to the public, leaving no room for further considerations related to the nature of the platform’s intervention or to its profit-making intention – considerations that the CJEU has found essential in recent case law14.

On the other hand, when the draft Directive was published there was no clear indication in the copyright acquis that the concept of communication to the public within the meaning of Article 3(1) InfoSoc Directive covered the activities of platforms of user-uploaded content. Arguably, some such platforms would nonetheless be found liable on grounds that they exerted an “indispensable intervention” in providing access to protected works15. This was made clearer in the recent Pirate Bay judgment, in which the CJEU held that an online search platform that indexes, categorizes, deletes and filters content may actually engage in acts of communication to the public16. But the nature of the intervention of BitTorrent websites like The Pirate Bay differs from that of online platforms that store and provide public access to user-uploaded content. This CJEU “precedent” could well be found inapplicable to platforms that do not show as high a degree of editorial intervention as The Pirate Bay17.

The Commission therefore (mis)uses Recital 38 of the draft Directive to reshape the concept of communication to the public, extending its coverage to new situations. As is common knowledge among jurists, in general, and among those interested in EU law matters, in particular, recitals are (supposed to be) interpretative tools. According to the Interinstitutional Agreement of 22 December 1998 on common guidelines for the drafting of Community legislation, “[t]he purpose of the recitals is to set out concise reasons for the chief provisions of the enacting terms, without reproducing or paraphrasing them. They shall not contain normative provisions or political exhortations”18. The CJEU has also acknowledged that recitals should not perform a normative a role, holding that “the preamble to a Community act has no binding legal force and cannot be relied on as a ground for derogating from the actual provisions of the act in question”19.

In agreeing that recitals should not assume the role of their act’s operative part, EU institutions show a palpable concern with legal certainty and with the protection of the expectations of legal subjects. In the draft Directive, the Commission grossly neglects these values, by including a crucial normative development of the copyright acquis in Recital 3820. Assuming such a development is to be adopted at all, it should be included in the operative part of the Directive and subject to the corresponding scrutiny.

1.2. The proposal lacks basic clarity

Legal norms, if they are to give their subjects any guidance, must be clear. “The desideratum of clarity”, in the words of a celebrated legal theorist, “represents one of the most essential ingredients of legality”21. The value gap proposal fails to achieve that desideratum. In fact, Article 13(1), read together with Recital 38, is a remarkably confusing provision22.

As seen, Recital 38 determines that online platforms that store and provide public access to protected content uploaded by their users perform acts of communication to the public. These platforms are thus obliged to conclude licensing agreements with rightholders, unless they are eligible for the safe harbour provided in Article 14 E-Commerce. However, platforms’ liability for communicating works to the public relates to the protection conferred by the safe harbour is not made entirely clear. Is the hosting safe harbour supposed to work as a defence


12 See Stall-a-Bourdillon et al. (2016a); Stall-a-Bourdillon et al. (2016b); Angelopoulos (2016); Angelopoulos (2017); European Copyright Society (2017); Max Planck Institute for Innovation and Competition (2017); and CREATe (2017). Scholars dispute the very existence of a “value gap”. See, e.g., Froiso (2017), p. 7 (arguing that the “value gap” rhetoric is “fabricated” by rightholders and “scarcely concerned with empirical evidence”), and CREATe (2017), p. 6 (claiming that “[t]he idea that the creation of value should lead automatically to transfer or compensation payments has no scientific basis”).

13 Such a codification of judge-made law would not be unprecedented. For instance, Article 4(2) of the InfoSoc Directive codified the rule of exhaustion of the distribution right developed by the CJEU in Deutsche Grammophon, C-78/70.

14 See CJEU, GS Media, C-160/15, paras. 47-55, and CJEU, Filmspeler, C-527/15, paras. 41-53.


16 See CJEU, The Pirate Bay, C-610/15, in particular para. 38. See also Rosati (2017), p. 15, noting that this recent judgment brought CJEU case law closer to the Commission’s legislative initiative.

17 See Angelopoulos (2017), p. 32.


19 CJEU, Nixson, C-162/97, para. 54.

20 In the words of Angelopoulos (2017), p. 31, “[f]or a nonchalant statement hidden in a recital, this represents a dramatic development of EU copyright law”.


22 See Angelopoulos (2016) and European Copyright Society (2017), p. 7, deeming the provision’s language “ambiguous”; and CREATe (2017), stating that the proposal is “poorly drafted”.

11 See Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Promoting a fair, efficient and competitive European copyright-based economy in the Digital Single Market, COM(2016) 592 final, p. 8. 12 See Stall-a-Bourdillon et al. (2016a); Stall-a-Bourdillon et al. (2016b); Angelopoulos (2016); Angelopoulos (2017); European Copyright Society (2017); Max Planck Institute for Innovation and Competition (2017); and CREATe (2017). Scholars dispute the very existence of a “value gap”. See, e.g., Froiso (2017), p. 7 (arguing that the “value gap” rhetoric is “fabricated” by rightholders and “scarcely concerned with empirical evidence”), and CREATe (2017), p. 6 (claiming that “[t]he idea that the creation of value should lead automatically to transfer or compensation payments has no scientific basis”). 13 Such a codification of judge-made law would not be unprecedented. For instance, Article 4(2) of the InfoSoc Directive codified the rule of exhaustion of the distribution right developed by the CJEU in Deutsche Grammophon, C-78/70.
against primary liability, in the sense that if the platform acts neutrally its liability for infringing the right of communication will be negated? Or is the platform's non-neutrality (i.e., its active role) a requirement for it to be considered as performing an act of communication? Who bears the burden of showing what?

In any case, proceeds the recital, regardless of whether platforms are protected by the safe harbour, they need to take measures to guarantee the protection of content, so as to ensure the functioning of licensing agreements... agreements which, as stated in the very same recital, protected platforms do not need to conclude. Imposing on platforms that are immunised by the hosting safe harbour an obligation to ensure the functioning of licensing agreements that they need not have concluded is, at best, puzzling. For the sake of analysis, however, let us consider the case of a platform that provides its service non-neutrally, being thus ineligible for the safe harbour, and that therefore undoubtedly needs to license the content it hosts (at least according to Recital 38).

In that case, Article 13(1) kicks in, imposing two alternative obligations upon the platform, in order to try and close the value gap: the platform should (a) take measures to ensure the functioning of agreements concluded with rightholders for the use of their content; or (b) take measures to prevent the availability on their services of content identified by rightholders through the cooperation with the service providers.

The first obligation - when interpreted in light of Recital 38, which uses similar language in its third paragraph ("ensure the functioning of any licensing agreement") - translates into an obligation to take appropriate and proportionate measures to ensure the protection of content, such as implementing "effective technologies". These technologies, judging by the second sentence of Article 13(1), Article 13(3) and Recital 39, can be no other than content recognition technologies.

Alternatively, platforms should take measures to prevent the availability on their services of content identified by rightholders. What are the precise measures that platforms should take? The suggestion that Article 13(1) makes, yet again, is that platforms use "effective content recognition technologies". The question then becomes: How, in practical terms, does preventing infringement (second obligation) differ from ensuring the protection of content (first obligation)? Are they truly alternative - and, by necessity, different - obligations? A combined reading of Recital 38 and Article 13(1) indicates that the two (supposedly alternative) obligations laid down in the latter are, in reality, one and the same obligation - and a rather worrying one: that online platforms implement content recognition technologies. But it may well be that the Commission intended to give online platforms a true choice. What that choice is, though, remains unclear.

In sum, as I hope to have successfully showed, several parts of Article 13(1) and Recital 38 lend themselves to importantly diverging interpretations. This lack of clarity is possibly the product of a difficult, compromise-ridden drafting process. Understandable as that is, it is safe to anticipate that the proposed formulation will result in different transpositions by Member States, thus fostering disharmony and legal uncertainty.

2.1. The proposal uses trivially vague language

Vague words are words that have borderline cases, i.e. cases in which "one just does not know whether to apply the expression or withhold it, and one's not knowing is not due to ignorance of the facts". Typical examples of vague words are gradable adjectives like "bald", "rich" and "mature". We all know people that are clearly bald and people that are clearly not bald, but we also know people that are borderline bald (I would count myself as one such person). Legislators tend to avoid employing these trivially vague terms. For instance, in establishing the legal capacity to vote, legislators set an age of majority, instead of stipulating that "mature citizens" are entitled to vote. And if legislators want to tax rich people more heavily, they usually do so by reference to a numerical level of income, instead of imposing a higher rate on "rich people".

Sometimes, however, vagueness in the law is useful or simply unavoidable. The most notable examples of vagueness in the law are cases of extravagant vagueness. Every jurist is familiar with examples of legal norms deploying this type of language, such as norms exempting the "fair use" of protected works from liability or norms requiring adherence to a standard of "reasonable care". In certain circumstances, the situations that ought to be covered by a norm are so diverse that vague language allowing for a multi-dimensional evaluation is needed. One scholar offers "child neglect" as an example of a situation where a myriad of factors need to be taken into account: "[y]ou just cannot stipulate that, say, leaving a child unattended for n hours would constitute neglect." In defining its subjective scope of application, Article 13(1) uses a vague adjective. The provision applies to information society service providers that store and provide to the public access to "large amounts" of content uploaded by their users. "Large", like "rich" or "mature", is a trivially vague word - the type of vagueness that the law normally tries to avoid. In most cases, instead of resorting to this sort of language, the law defines a more or less precise threshold for its application. The exercise of defining such a threshold is, of course, fairly arbitrary; but there are very good reasons for doing it, namely legal certainty and efficiency.

While in some cases the flexibility provided by vague language is needed, the situation regulated by Article 13 is far from being one such case. If online platforms are to comply with an obligation to implement certain exacting measures, they need a high degree of guidance. They need to know, at the very least, if they are subject to that obligation or not. It is beyond doubt that YouTube and Dailymotion are platforms hosting "large amounts" of user-uploaded content. But what should be said about the Portuguese platform Sapo Vídeos or the German tape.tv? Do these platforms host large or non-large amounts of content? Where and how should the line be drawn?

2. Substantive problems

2.1. The proposal throttles digital innovation

By now, you hopefully agree that the value gap proposal is not an example of good legal drafting. Still, you may be tempted to downplay the criticism: after all, form is not substance and substance trumps form. On the substantive front, however, the value gap proposal does not fare any better.

As seen, in what regards the measures for copyright protection that should be adopted by platforms hosting large amounts of user-uploaded content, the only suggestion given by Article 13(1) and the cited recitals is the implementation of content recognition technologies.

This suggestion, if transposed by Member States as a mandatory requirement, has the potential to seriously threaten innovation in the digital economy. Requiring online platforms to use such technologies entails erecting a market entry barrier that is very costly to overcome. Google, for example, reported that, by 2016, it had invested more than $60 million on its ContentID system. An obligation to filter user uploads would discourage investment in the development of this type of platforms, with indirect adverse effects on user creativity.

The fact that the requirement may be imposed only upon hosts of "large amounts" of content does not eliminate the problem. At a certain point in their growth (a point that Article 13 leaves indeterminate), smaller platforms will be forced to invest in technology the development and
2.2. The proposal disproportionately restricts fundamental rights

Several commentators have shown a concern with the proposal’s compatibility with Article 15 E-Commerce, which establishes a prohibition on general monitoring obligations. In fact, the use of content recognition technologies necessarily involves such monitoring. And, again, the fact that the proposal applies only to some platforms is not enough to save it: the covered platforms, however few they may be, will have to monitor all of the new content that is uploaded onto them. Christina Angelopoulos puts the point metaphorically: “[t]he chaff cannot be separated from the grain without the threshing of all the harvested wheat”.

But the proposal’s incompatibility with the E-Commerce Directive is not, in and of itself, a decisive argument against the requirement of using content recognition technologies. A conflict between Article 13 of the Draft Directive and Article 15 E-Commerce would be solved by the well-known meta-rue lex posterior derogat legi priori. EU institutions may well want to amend the acquis, in which case the later rule, introducing an exception to the general rule prohibiting monitoring obligations, would prevail.

But the commentators’ concern runs deeper than this. The conflict with Article 15 is particularly worrying because it is, simultaneously, a conflict with the Charter of Fundamental Rights of the EU. As hinted by the CJEU, the prohibition against general monitoring obligations is rooted in Articles 8 (protection of personal data), 11 (freedom of expression and information) and 16 (freedom to conduct a business) of the Charter.

In the name of safeguarding the interests of rightholders, an obligation to implement content recognition technologies would disproportionately restrict the Internet users’ right to the protection of their personal data, as well as their freedom of expression and information. First, content recognition technologies would necessarily involve the “identification, systematic analysis and processing of information” connected with the profiles of individual users, allowing them to be identified. Second, these technologies are not infallible: they often fail to adequately distinguish between lawful and unlawful content, taking down uses that may be protected by copyright exceptions. Not long ago, it must be noted, the European Parliament invoked these rights when rejecting the adoption at the EU level of enforcement strategies based on a three-strikes policy (the so-called “graduated response systems”).

In addition, as suggested above, the obligation to use content recognition technologies would gravely affect the platform operators’ freedom to conduct their business, since it would require them “to install a complicated, costly, permanent computer system at [their] own expense”.

Contrary to what is stated in its Recital 45, the draft Directive does not respect the fundamental rights enshrined in the Charter. The value gap proposal fails to strike a fair balance between, on the one hand, (i) the users’ right to the protection of their personal data, (ii) the users’ freedom of expression and information, and (iii) the online platform operators’ freedom to conduct their business, and, on the other hand, the intellectual property rights of creators, protected by Article 17(2) of the Charter. If enacted without the necessary amendments, Article 13 may be struck down by the CJEU on grounds of violations of the foregoing rights.

Conclusion

The value gap proposal, as provided in Article 13 and Recitals 37 to 39 of the draft Directive, suffers from a number of fundamental problems. From the formal point of view, the proposal (i) includes normative provisions in the recitals, (ii) lacks basic clarity, and (iii) uses trivially vague language. On the substantive side, if passed into law the proposal will (i) thwart digital innovation and (ii) disproportionately restrict the fundamental rights of Internet users and platform operators.

Given the seriousness of these flaws, the EU institutions should consider deleting or, at least, significantly rewriting the proposal. The amendments put forward by former rapporteur Therese Comodini Cachia address some of the mentioned flaws and could form the basis of a possible rewrite. Nevertheless, the amended provisions fail to give targeted platforms guidance regarding the appropriate and proportionate measures they should take, while preserving a risky reference to content recognition technologies in Recital 39. At this stage, and considering the extent of the necessary amendments, it should be kept in mind that it may be better to pass no text than to pass a text that promotes legal uncertainty and threatens innovation in the digital economy.

32 See infra 2.2.
33 See Angelopoulos (2017), pp. 37-38. See also Husovec and Leenes (2016), pp. 47-48, reporting on the costs behind the implementation of filtering technologies.
37 In any case, such a change should be made clear in the enacting terms of the Directive, rather than left tacit.
38 See CJEU, Scarlet, C-70/10, paras. 46-50, and CJEU, Netlog, C-360/10, paras. 44-48.
39 See CJEU, Scarlet, C-70/10, para. 51, and CJEU, Netlog, C-360/10, para. 49.
40 See CJEU, Scarlet, C-70/10, para. 52, and CJEU, Netlog, C-360/10, para. 50.
41 See European Parliament, Resolution of 10 March 2010 on the transparency and state of play of the ACTA negotiations, para. 11.
42 See CJEU, Scarlet, C-70/10, para. 48, and CJEU, Netlog, C-360/10, para. 46.
43 Although the CJEU has been assessing the validity of EU legislation against fundamental rights for some time now, it has been more open to engaging in such close scrutiny since the Charter became legally binding. See, notably, CJEU, Digital Rights Ireland C-293/12, in which the CJEU annulled the Data Retention Directive, on grounds of violation of Articles 7 and 8 of the Charter.
References


Fuller L (1964) The Morality of Law. Yale University Press, New Haven


Kretschmer M et al. (2016) The European Commission’s Public Consultation on the Role of Publishers in the Copyright Value Chain: A Response by the European Copyright Society. EIPR 38(10):591


Ramalho A (2017) Beyond the Cover Story – An Enquiry into the EU Competence to Introduce a Right for Publishers. IIC 48(1):71


Copyright on Data
**Introduction**

On 10 January 2017, the European Commission issued its Data Economy Package, which includes a Communication on building the European data economy\(^1\) and an accompanying Commission Staff Working Document on the free flow of data and emerging issues of the European Data Economy,\(^2\) followed by a Public Consultation. The initiative reveals the growing importance of non-personal, machine generated data. It is grounded on the need for market players to have access to large and diverse datasets in the context of the emergence of a data economy, as well as on the goals of incentivizing the sharing of data, ensuring the free flow of data, protecting investments and assets, and minimizing lock-in effects.\(^3\) The number of market players involved in data trading as a means to develop new business models or to open up additional revenue streams is expected to grow exponentially.\(^4\)

Personal data is excluded from the scope of the communication,\(^5\) even though the Commission recognises that some market players deal with datasets that contain both personal and non-personal data\(^6\) (this would be the case, for example, of wearables that function as health and fitness trackers). The data under consideration in the Communication are both non-personal (either naturally non-personal, or turned non-personal through the process of anonymization) and machine-generated (i.e., created without human intervention, through e.g. computer processes or applications).\(^7\) Arguably, the exchange and access to this type of data is limited, with many companies that de facto own the data generated by their products or services usually preferring not to share it, and, according to the Commission, even preventing the user who owns the data-generating device from authorising use of the data by other companies.\(^8\)

The Commission puts forth several possibilities to increase access to and sharing of data, such as e.g. setting up default data contract rules, fostering the development of technical solutions for reliable identification and exchange of data, developing a framework based on FRAND (fair, reasonable and non-discriminatory) principles to provide access to data against remuneration, or creating a data producer’s right. This contribution will analyse the latter solution. As the Commission rightly points out, raw machine-generated data are not protected by any intellectual property rights, and their economic exploitation and exchange is frequently ruled by contract.\(^9\) Indeed, typically, IP rights are granted at the innovation (or expressed creativity) level. By contrast, the production of data happens at an earlier stage in the data value chain, prior to any innovation.\(^10\)

---

\(^1\) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, “Building a European Data Economy”, COM(2017) 9 final, 10.1.2017 [hereinafter, “Communication”]


\(^3\) Communication, at 4.


\(^5\) Legislative initiatives concerning personal data have already been undertaken separately, see e.g. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), or the review of the e-Privacy Directive (which resulted in a proposal for a Regulation concerning the respect for private life and the protection of personal data in electronic communications, COM(2017) 10 final, 10.1.2017)

\(^6\) Communication, at 9.

\(^7\) Communication, at 9.

\(^8\) Communication, at 9-10. See also SWD, at 15-16, where the Commission points to a notable exception to the alleged non-sharing trend: the bank BBVA shares anonymized and aggregated statistical data from millions of transactions that can convey consumer’s habits or demographics.

\(^9\) Communication, at 10.

The question therefore becomes whether the creation of a data producer’s right is an adequate solution. The next section will briefly describe the contours of the proposed right. The following sections will focus on the merits of the right and analyse respectively its powers, perils and pitfalls.

**The data producer’s right**

As devised in the Commission’s Communication, the subject matter of the data producer’s right is non-personal or anonymised machine generated data, including metadata on the data. The Commission stresses that the data covered by the right should be at the syntactical, not the semantic, level, and that care should be taken to ensure that ideas and information remain free.11 This means that the object of protection is at the level of signs (such as sequences of 0 and 1), not at the level of content of the information.12

The owner of the right would be the data producer (i.e., the owner or long-term user of the device), who would then have the exclusive right to use and authorise the use of the data (e.g. through licensing). According to the Commission, this would include “a set of rights enforceable against any party independent of contractual relations thus preventing further use of data by third parties who have no right to use the data.”13 The right would be limited by exceptions granting access to the data by others, namely the manufacturer of the device (who, besides having a commercial interest in the data, might be obliged by national law to monitor the product) or public authorities (for, e.g., statistical information or urban planning).14 The Commission further envisages that, in certain cases, there might be a public interest in making the data available for other private actors, such as sharing smart metering information for purposes of fully enabling smart homes or care institutions.15 Along the same lines, an exception to the right could also be established to ensure access for research that is entirely or mostly funded by public resources.16

The right addresses a controvert question in the data economy – who owns the data? As pointed out by the Commission, given the regulatory gap in this regard, the (de facto) owner of the data is the company whose devices generate the data. The data producer’s right would shift the (de facto) data ownership from the company that supplies the machines or devices to the user/owner of the device, allowing the latter to contract with other data-based service- and device providers. This is connected to the objectives of the right, which are “clarifying the legal situation and giving more choice to the data producer, by opening up the possibility for users to utilise their data and thereby contribute to unlocking machine generated data.”17 Presumably, clarifying the legal situation by attributing exclusive rights would avoid conflicts over ownership, and giving more choice to the producer would contribute to fostering data access and sharing (and thereby the data economy).

**Powers**

Powers have largely been overlooked until now. The objectives of the Commission’s proposals are laudable and include the improvement of access to non-personal machine-generated data, the facilitation of its sharing, and the protection of investment that also takes into account a fair sharing of benefits with other players.18

The Commission is also well aware of potential risks that the current situation might bring to the development of a sound data market. In fact, it should be noted that, in most cases, manufacturers or service providers de facto own the data, which could in theory lead to unfair contractual terms of access to the data. The problem of access to data is – rightly – at the centre of the Commission’s line of action,19 and even if the data producer’s right is in the end an exclusive right over data, it is also mainly thought of as a way to ensure access, rather than as a way to generate income from further uses of the data.20

**Perils**

Even though the intentions of creating a new data producer’s right might be commendable, such right would also have its shortcomings. Perils or immediate dangers include the fact that creating a new right will add an extra layer of rights to be cleared, which in turn can work against one of the other objectives of building a data economy – the free movement of data.21 Moreover, the right becomes especially problematic if one considers that non-personal data becomes most valuable when used in large amounts (big data). Giving exclusive rights over small amounts of data will hinder big data analytics (since the analysis would require acquiring lots of exclusive rights held by different owners), with potentially negative effects to the data economy.

Furthermore, in practice, the manufacturer of the device could just resort to contracts to regain control of the data (e.g. through an exclusive license), in which case de facto control becomes legal control. Put it differently, an IP right would not solve the problem of the de facto control, since the manufacturer of the device will typically be in a stronger negotiating position and can contractually acquire the rights (a better solution for a situation of de facto control could be, e.g., competition law).22

Yet another peril is the risk of information lock-ins due to the difficulty in distinguishing the syntactical from the semantic level. The value of the data comes from the information it can convey, and from the insights that can be derived from it (i.e., the semantic level); but the information at the semantic level can be transformed into data (at the syntactic level), which means that protection of one can entail protection of the other.23

**Pitfalls**

**Mismatch between IP rationales and the data producer’s right**

The first pitfall, or source of potential danger, is the mismatch between justifications or rationales for IP protection, on the one hand, and both the subject matter of the data producer’s right (machine generated, non-personal data) and the objectives underlying the protection of non-personal data, on the other hand. This mismatch could dictate the inadequacy of the right to achieve the goals it is supposed to achieve.

---

11 SWD, at 34.
13 SWD at 33.
14 SWD at 35-36.
15 Ibid.
16 Ibid.
17 Communication at 13.
18 Communication, at 11.
19 Stakeholders have also highlighted that the main issue is access: see SWD at 35.
21 As pointed out by the Commission in the Communication, at 3, 5 et seq.
Several theories compete regarding the justifications or rationales for intellectual property rights. Each of them is conducive to a different analysis of the same IP policy, but new exclusive rights should be viewed from the perspective of all of them before they are introduced in the legal order.

The personality rights theory, for which mainly the philosophers Kant and Hegel are responsible, holds that the work is worthy of protection because it is an expression of the personality or self of its creator.25 Not all types of intellectual property can be justified by this theory; rather, only those that entail some level of personality or self-expression (such as copyright).26 This connection to the personality of the creator is at odds with the very nature of the subject-matter of the data producer’s right (non-personal data).

The labour theory, formulated by the British philosopher John Locke in the 17th century, implies that every man should be the proprietor of the product of his labour. This theory suggests the idea of “reward”, as it would seem fair that whoever uses his or her intellectual labour to create or to invent should have a right over the ensuing product.27 Non-personal, machine generated data does not fit this construction, as there is no intellectual labour in the production of such data. There might be intellectual labour involved in creating devices and sensors for the production of data – and those devices and sensors could be protected by an IP right if conditions are met – but not in the further production of data itself.

Finally, the utilitarian theory considers that IP rights are necessary to promote or incentivize creative/inventive activity,28 but also the efficient use and commercialization of the creation or invention after it has been produced.29 Under this theory, IP rights are positive rights (as opposed to natural rights), granted with the goal of furthering societal welfare. In other words, IP rights are granted to creators and inventors with the goal of promoting further creation and inventive activities, and the dissemination of the outcomes from said activities (namely, through licensing). It is doubtful that the utilitarian rationale, based on the idea of incentive, can justify the creation of an IP right in data. Data is produced independently of incentives; there is no underproduction of data that needs to be remedied (in fact, quite the contrary is true).30 It is not because of the existence of an exclusive right over data that users of devices that generate data will start producing more data. Likewise, a new exclusive right will also not facilitate access or sharing, as firms are able to trade data without having exclusive rights on it (mainly relying on their de facto control).31 The incentive rationale for the creation of an IP right in data is thus not present either.

It should be noted however that other rights, such as the sui generis right for databases or some neighbouring rights, have a slightly different justification – the protection and/or promotion of investment.32 It is the case, for instance, of the neighbouring right of the film producer33 or the sui generis right of the database maker.34 It is doubtful that a data producer’s right could be devised as a neighbouring right as such. Neighbouring rights are an umbrella category that encompasses rather different subject-matter, but generally they aim to protect activities somehow related to copyright. Their object is the dissemination – not the creation – of works which are often literary and artistic works.35 Raw data which is machine-generated does not fall under copyright, which means that the link between a (possible) neighbouring right and copyright is not present in the case of a data producer’s right. This makes it difficult to justify the use of a neighbouring right for the case of non-personal machine generated data.

The last possibility available would be to have a sui generis right for data producers, which would have as its main justification the protection and/or promotion of investment. However, if the data producer’s right is designed as a “one size fits all”, it should also be considered whether all investments (from owners or long-term users of devices) are worth protecting. This problem is connected to another issue: as the Commission rightly notes, it is hard to identify clear patterns across different sectors, with usage rights being dependent on context and the particular situation at stake.36 The take-up – both legal and commercial – also plays a role when it comes to the importance of non-personal data to the data producer. Arguably (and on average), the consumer who generates non-personal data through his or her fitness tracker will be less interested in questions of (non-personal) data ownership than the owner of a smart factory’s machinery; the former will also have less need in acquiring the device/machine than the latter. This means that a legislative solution designed as a “one size fits all” measure may in any case be inadequate and premature, regardless of its legal form. Moreover, most of the stakeholders that participated in a study about the European Data Market are satisfied with current arrangements (mainly contractual) for the exchange of data, conveying that the current levels of data exchange and re-use do not seem to cause problems to the market efficiency.37 In its Staff Working Document, the Commission refers to the need to define rights of access to data than to define ownership rights.38 Altogether, it is doubtful that there is an investment worth protecting in every case where a data producer’s right would be applicable.

Moreover, IP rights, including neighbouring rights or the sui generis right of databases, share one common trait – they stem from human creativity or effort (be that effort financial or intellectual), to a greater or lesser degree. Machine-generated data seems to be one step further than that – it can be generated automatically, without any human intervention (apart from the

32 This is true for some neighbouring rights – such as the rights of producers – but not for others – such as the rights of performers. Protection for the latter is based on social objectives and natural rights arguments, as well as on utilitarian arguments – see M. van Eechoud et al., Harmonizing European Copyright Law. The Challenges of Better Lawmaking, Wolters Kluwer, 2009, at 186-194.
33 Recital 5 of Rental Right Directive (Directive 2006/115/EC of the European Parliament and of the Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intelectual property, O.J. L 376, 22 December 2006) notes that there is no investment worth protecting in every case where a data producer’s right would be applicable.
34 SWD, at 16. Also pointing out the problems with the “one size fits all” approach, J. Drexl, “Designing Competitive Markets for Industrial Data – Between Propertisation and Access”, at 40.
35 35
36 SWD, at 16. Also pointing out the problems with the “one size fits all” approach, J. Drexl, “Designing Competitive Markets for Industrial Data – Between Propertisation and Access”, at 40.
38 SWD at 35.
The creation of data left outside the scope of the sui generis right includes principle data generated by machines. In any case, at least in theory, many businesses could have (an aspect of) non-personal data protected via the Database Directive if they so wished, because e.g. they invest in creating and managing the contents of an app. Admittedly, however, this is not in the interest of all de facto owners of data, with some preferring to keep the data to themselves (for purposes of product improvement, for instance). More importantly, it can be difficult to differentiate investment in creating or producing the data (for example, through sensors) and investment in obtaining or collecting it41 (which in the case of machine generated data could be, e.g., assembling the data from several devices into files). Where it is not possible to distinguish between these investments, obtaining the raw data, or where the activity of obtaining the data is “indivisibly linked” to the creation of data, the CJEU has considered that there is no independent investment in obtaining the data (thus denying protection to the database on those grounds). 42

Moreover, depending on how a potential data producer’s right is designed, conflicts might arise due to the fact that the database and the machine-generated data (which might later go in the database) have different owners: the owner of the former will be the database producer, while the owner of the latter will be the user who owns or is in possession of the device. Even though the manufacturer would have a non-exclusive access to the data, determining potential infringement by third parties could be challenging.

**Drawing the line between personal and non-personal data**

In some devices such as wearables, personal and non-personal data are intertwined, and it might be difficult to draw the line between them. Importantly, such line must be drawn, as personal data is subject to its own specific regime in the EU. Personal data concerns information where a natural person is identified or identifiable, including personal data that have undergone pseudonymization but that could be attributed to a natural person by using additional information.43 In the assessment of whether a person is identifiable, account should be taken of the means used to do so, such as singling out, including its cost and amount of time required for identification.44 According to Article 4(1) GDPR, a natural person is identifiable by reference to “a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”. The CJEU has added to this list an IP address.45 The Court has also stressed that, in order to treat information as personal data, it is not necessary that that information alone does not identify the subject to be identified.46 The definition of “personal data” is thus by no means straightforward. In addition, the nature of data is dynamic and subject to change. Anonymous data can be deanonymized, for example, by matching it with other datasets and applying some probability theory.47 This makes the distinction between personal and non-personal data a moving target.

**Conclusion**

Enhancing data sharing and access and doing away with legal uncertainty in data markets are in themselves praiseworthy objectives. The solution of achieving them through introducing in the legal order a new property, IP-type of right is however not the best course of action.

---

48 By Velkkau, at 44; Svenska Spel at 33; ORAP at 49.
49 Recital 26 of the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679 of the European Parlia- ment and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of per- sonal data and on the free movement of such data, and repealing Directive 95/46/EC), O.J. L119, 4 May 2016. 50 Ibid.
51 In case C-582/14 #breyer, the CJEU has considered that in cases where the IP address is capable of sufficiently identifying a natural person (because the provider has means to identify the person with additional data) such address amounts to “personal data” (see para. 49 of the decision).
52 See however, an exception to this, Section 9(3) of the Copyright Designs and Patents Act of the United King- dom, which vests protection of computer generated works in the person by whom the arrangements necessary for the creation of the work are undertaken.
40 Communication at 12-13.
41 A proportionality check includes the evaluation of three factors: the suitability of the measure for the attainment of the objective; the necessity of the measure; and the proportionality of it vis-à-vis the restrictions that might be thereby involved, or proportionality strictu sensu. See inter alia case C-331/87 Fedesa, para. 13, and case C-210/00 Alseni et al. v. Compagnie Champignon Hofmeister, paras. 59-67. For a detailed explanation of the factors, see X. Grousset, General Principles of Commu- nity Law, Europa Law Publishing, 2006, at 146-152; J.H. Jans, “Proportionality Revisited”, Legal issues of Economic Integration 2000, 27(3), at 240 et seq., and references therein.
45 Recital 46 Database Directive.
46 Case C-46/02 Oy Velkkau, at 14; case C-203/02 British Horseracing Board at 31; case C-338/02 Svenska Spel, at 24; case C-444/02 ORAP, at 40.
49 Recital 46 Database Directive.
46 Case C-46/02 Oy Velkkau, at 14; case C-203/02 British Horseracing Board at 31; case C-338/02 Svenska Spel, at 24; case C-444/02 ORAP, at 40.
The fabric of a data economy is not compatible with exclusive, crystallised IP rights: a data producer’s right would add an extra layer of rights to the legal order, which could hinder the free flow of data; rights over small datasets would be at odds with the big data analysis that underlies the data economy; and exclusive rights over data could also lead to information lock-ins. None of the justifications for having IP rights are fulfilled in the case of a right over data. There is no connection to the personality of the creator, nor to its intellectual labour. No incentives are needed to produce or disseminate data. Even the case for a sui generis right is weak, since there is no investment worthy of protection or promotion (at least not in all cases where a data producer’s right would be applicable).

The interaction of the data producer’s right with existing regimes, namely the sui generis right for databases and the protection of personal data, could also lead to conflicts and result in legal uncertainty.

The objectives of the Commission would thus be better achieved through other options mentioned in its Communication, which sound both more efficient and realistic. Such is the case, for instance, of the implementation of an obligation to license the re-use of data under fair, reasonable and non-discriminatory terms (FRAND), or developing guidelines to incentivise businesses for the results of mining. If TDM is considered to be encouraged for social benefit, policy makers might therefore take into consideration these various needs and eventually intervene in order to build the necessary foundations in case the market does not provide for – investing in digitization, establishing norms for formats, facilitating open or broad availability of the content and the technical tools, developing storage, cloud computing facilities… It is also noticeable that access to information is not always locked by IP rightholders but may be the mere results of the technical tools, allowing the process (the picks and shovels);

The first invariable criterion is the digital format of the source subject to the process of mining, the second is the existence of a process involving specific tools, the third is the purpose for which the process of the element is being made (animus), the fourth is the constitution of a result of the process that is new/different from the source mined.

In order to actually realize TDM, it is necessary to have access to the “source” (the mine) and to the tools allowing the process (the picks and shovels); to search something and to have a reasonable expectation of the results that may derive from the searching activity (the opportunities of exploitation of the ore).

Many of the elements required to achieve TDM are irrespective of copyright issues such as the need for technical infrastructure and investment allowing the miners to dig, the market opportunities for the results of mining. If TDM is considered to be encouraged for social benefit, policy makers might therefore take into consideration these various needs and eventually intervene in order to build the necessary foundations in case the market does not provide for – investing in digitization, establishing norms for formats, facilitating open or broad availability of the content and the technical tools, developing storage, cloud computing facilities… It is also noticeable that access to information is not always locked by IP rightholders but may be the mere results of contractual practices and/or technical control of access of individual or companies regardless of any IP consideration.

Therefore, solving the copyright issues that may conflict with TDM will not be the only key for suddenly developing the market of the applications of TDM in Europe, this will also be a matter of education, investment, interoperability, open data policy. Some publisher’s opponents to the TDM exception point out that they don’t face an important demand for TDM license when required (hardly few licenses a year) and that they easily come to an agreement in this case. Many reasons may explain why they experience such a situation (people don’t ask permission because they don’t know they have to, they know but they fear to
Part II. Which are the problems with mining vis-à-vis copyright issues?

Going back to the “bony” definition of TDM, TDM uses may encroach on Intellectual property rights. The “source” of the TDM might be protected by different rights – copyright, neighbouring rights, sui generis rights on databases-. So access to the source may – in certain circumstances – trigger the application of IP rules.

The type of tool used to mine may also raise questions of IP when it comes to the acts of exploitation involved in the process or compliance thereof with the digital right management.

The assessment of the purpose/intent may also be taken into account when considering the balance between the claim for exclusive right on the one hand and the claim for accessing to and processing the content. At this stage, this shall be limited to a conflict of principles- who is entitled to mine with which project against the legitimacy of the rightholder position.

Finally, the opportunities of “exploitation” of the result of the mining may also be balanced according to the competing interest of the rightholder to benefit from its property.

2.1. Mining a “protected” material

This paper will not detail the famous distinction between the “form” of expression that is the subject protected by copyright rules and the mere information or ideas that are outside of the scope. We shall only insist on the fact that the distinction is legally and practically fragile. Even if one might find some traces of the distinction in the international and European “acquis” such as the article 9.2 of TRIPS agreement or article 2 WCT and more specifically in the computer program directive article 1.2 (Protection in accordance with this Directive shall apply to the expression of a work in any form of a computer program. Ideas and principles which underlie any element of a computer program, including those which enable its interfaces, are not protected by copyright under this Directive.)

it is still complicated to draw a line between the form and the information whenever access to information supposes reproduction of the form in which this information has been expressed.

2.2. Mining as an act covered by IP rights

Besides, many copyright rules do cover the use of the information/idea/content. The value of the work is also linked to the amount of information it provides to the public – see the newspaper- the content is indirectly covered by the copyright protection when an authorization is required for the adaptation of a novel into the movie: it is the story, the characters that matter here and not the choice of a specific sentence. When considering moral right to integrity, distorting the “spirit” of the work may amount to an infringement before certain jurisdictions. Furthermore, the formation–dissimulation dilemma has less echo in the realm of neighboring rights and sui generis rights for databases. Even if the 1992 directive remain silent on the definition of the subject-matter (phonogram, film...), of the producer or performer rights the ECJ and the recitals of the various directives seem to acknowledge that the justification for protection lies in the investment made by the producer, just as the maker of the databases for the sui generis right, whereas the performers rights are covering the performance and its fixation. If access to the data contained in this “material” supposes any process of copying of the file, this might trigger the protection because investment has been protected. According to the rules governing the relationship between copyright and related rights, it would be paradoxical to consider that the use of the data in the work is free because of the form/information dichotomy whilst the use of the data in the related right is not. If TDM is to be enhanced because of its social function, it is questionable to dissociate its regime according to the various regimes of the sources.

Yet, if we stick to mere copyright consideration, it could be argued in order to limit the conflict between the claim for free use and the exclusive right that TDM does not access to and/or use a “work” per se, according to a “functional” conception of what a work is. For example, in France some decisions (see example in France: Etre et Avoir / Place des Terreaux) have considered that there was no reproduction of the work “as such” when the public could see it (architectural work or drawing) in the frame of a wider image or in a film but that the work was not the subject of this image. Instead of relying on a legal exception to the exclusive right (namely panorama and ancillary copy that was not existing in the law) the judges answered at the upper level and decided that the work as a legal concept was not sufficiently “present”; therefore, that there was no infringement of the copyright. It is the metaphor of the puzzle: even if all the pieces of the puzzle are present, the recombination of all of them do not amount to the same source.

Consequently, one of the first answer that could mitigate the conflict between TDM and copyright would consist in defining the works subject to protection in the acquis – when this definition only exists for computer programs, databases – and/or determining the situations when the work would not be protected because the function of copyright has not been harmed. Accordingly, all the works would not be protected and some could be explicitly excluded from the scope of copyright and the protection of the work would not be absolute but may depend on the function provided for in the copyright legislation. Still even in introducing in the directive the distinction between legal and physical expressions and information and protection is limited to work “as such” it is rather unsure that it would provide enough legal security and fulfill the objective of harmonization of the situations within the different Member States.

One can also think of relying on the concept of originality and consider that the threshold of protection is not met in certain cases. This question has been addressed by the Court of Justice in the famous Infopaq case (rendered about clipping practice of the press aggregators) and the judges held that reproduction of small pieces of the work – extracts of 11 words – was to be considered according to the originality criterion. The decision has been interpreted as acknowledging the originality of the pieces but I think that we can also see it the other way round like establishing the possibility that the originality threshold is not sufficiently fulfilled in certain cases and that the judges have to control the existence of the originality of the “pieces”. It would not be sufficient for the rightholder to prove that the source as a whole was original to prevent the use of the pieces, since it should also be demonstrated in this case that the pieces are also – as such – original. So legislative intervention could consist in defining a general threshold of originality.

Still, even if defining what a “protected work” means as regards EU law would certainly be needed in the broader perspective of a coherent copyright code, and may help to delineate when a “work” is used, it does not seem to match the short-term perspective of the ongoing harmonization.

2.3. Mining as an act covered by IR rights

Besides, as regards TDM, the difficulty lies in the fact that the process involves technical reproduction of works and that the Infosoc Directive cover any kind of reproduction by an exclusive right (art. 2) (Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part). This very broad definition of the reproduction right has been heavily criticized by many authors in France because our definition (still) requires an act of communication to the public of the work (Article L. 122-3 Code de la propriété intellectuelle: Reproduction shall consist in the physical fixation of a work by any process permitting it to be communicated to the public in an indirect way.)
According to this requirement, it can be argued that some acts of reproduction are not covered by the definition of the exclusive right since they don’t permit to the communication to the public of the work. What is communicating the work to the public? I don’t refer here to the flimsy definition provided for by the Court of Justice in its numerous case-law but to a “sensitive” perception of what is an act of communication towards human being. Is a work communicated to the public when the recipient cannot perceive, recognize it? When considering copyright infringement, the judge compares the resemblances between the work and the copy: it is a “human” appraisal of where the work has been reproduced. So keeping a broad definition of the right of reproduction covering any kind of reproduction – even individual, temporary, transient, partial, automatic – may, at the end of the day, be inconsistent with the assessment of the infringement.

So an answer to TDM but also to other issues (linking, transient copies) may consist in introducing a condition of human identification of the work. Copyright protection would only be involved as far as the work is perceptible, recognizable by a human (and not a mere machine): the material presence of the work would not be sufficient to infringe copyright if this presence is not somehow perceptible by the public. Article 5.1. of the Infosoc directive – establishing the transient exception relies somehow on this assumption of human perceptibility when saying that the reproduction which is necessary to the transmission of work in a network is “exempted” from the reproduction right. But instead of establishing positively the condition in the definition of the exclusive right, it is in the rationales for a mandatory exception.

Being a pragmatic person, and though I regret it, I don’t believe that the legislator will have the courage to introduce such a “perceptibility” condition in the definition of the exclusive rights but will overcome the difficulty with exceptions. This requirement would indeed - in the present context of value gap - deprive the rightholders from the possibility to be associated to the value deriving from the automatic processing of their work, which appears to be a very promising market. Yet, it is my belief that mass digitization of works - whatever the purpose is: linking, mining, crawling - implies other answers than the mere individual exclusive right and that establishing a differentiated regime of protection depending on the existence of a “sensitive” contact of the human being with a work at the end of the process would be a solution.

As to TDM, art 5.1. provides a part of the answer and may cover most of the reproduction acts involved in the process of mining. This appear in the directive proposal, where it is recalled that the reproduction which is necessary to the transmission of work in a network is “exempted” from the reproduction right. However, the mining, crawling - implies other answers than the mere individual exclusive right and that establishing a differentiated regime of protection depending on the existence of a “sensitive” contact of the human being with a work at the end of the process would be a solution.

To TDM, art. 5.1. provides a part of the answer and may cover most of the reproduction acts involved in the process of mining. This appear in the directive proposal, where it is recalled that the reproduction which is necessary to the transmission of work in a network is “exempted” from the reproduction right. However, the mining, crawling - implies other answers than the mere individual exclusive right and that establishing a differentiated regime of protection depending on the existence of a “sensitive” contact of the human being with a work at the end of the process would be a solution.

To TDM, art. 5.1. provides a part of the answer and may cover most of the reproduction acts involved in the process of mining. This appear in the directive proposal, where it is recalled that the reproduction which is necessary to the transmission of work in a network is “exempted” from the reproduction right. However, the mining, crawling - implies other answers than the mere individual exclusive right and that establishing a differentiated regime of protection depending on the existence of a “sensitive” contact of the human being with a work at the end of the process would be a solution.
Data Property: Unwelcome Guest in the House of IP

P. Bernt Hugenholtz
Institute for Information Law (IViR), hugenholtz@uva.nl


1. Introduction

With the incessant growth of the ‘data-driven economy’ have come calls for the introduction of a novel property right in data. Apparently in response to demands from the automotive industry, and encouraged by a number of German lawyers and scholars, the European Commission has in its 2017 Communication on ‘Building a European data economy’ tentatively advanced the idea of creating at EU level a ‘data producer’s right’ that would protect industrial data against the world. The movement for ‘data property’ (in German Dateneigentum) has its champion in European Commissioner Günther Oettinger, who until 2016 led the directorate general that is responsible for the Communication, DG Connect. An op-ed published by Mr. Oettinger in the Frankfurter Allgemeine Zeitung reveals some of the thinking and the powerful forces behind this revolutionary legal concept. Data, writes Oettinger, are the “gold of the future”, principally in the automotive sector where modern sensor-equipped cars automatically generate and collect large amounts of data – on traffic and road conditions, engine performance, etc. These machine-generated sensor data have enormous value, for example, for developing self-driving automobiles. But – writes Oettinger – it is as yet unclear who owns these data: the automobile manufacturer; the car owner; the producer of the sensor equipment; or no one at all? What we need, concludes the Commissioner, are rules at EU level that establish data ownership. Apparently inspiring this call for protecting industrial data is the fear – common to other recent policy initiatives – that valuable European assets are being misappropriated by large American companies. The specter of Google ‘stealing’ European news has already led to an ongoing EU initiative towards a neighbouring right for news publishers, following comparable rules previously introduced in Germany and Spain. The sui generis database producer’s right introduced in Europe in 1996 was similarly inspired by European fears of dominance by the US database industry.

Although the contours of the ‘data producer’s right’ being contemplated by the European Commission are sketchy, as are its economic underpinnings, such a right would most likely bring the protection of industrial data in the EU to a much higher level than the – much-maligned and still unimplemented – sui generis database producer’s right introduced in Germany and Spain. The movement for ‘data property’ (in German Dateneigentum) has its champion in European Commissioner Günther Oettinger, who until 2016 led the directorate general that is responsible for the Communication, DG Connect. An op-ed published by Mr. Oettinger in the Frankfurter Allgemeine Zeitung reveals some of the thinking and the powerful forces behind this revolutionary legal concept. Data, writes Oettinger, are the “gold of the future”, principally in the automotive sector where modern sensor-equipped cars automatically generate and collect large amounts of data – on traffic and road conditions, engine performance, etc. These machine-generated sensor data have enormous value, for example, for developing self-driving automobiles. But – writes Oettinger – it is as yet unclear who owns these data: the automobile manufacturer; the car owner; the producer of the sensor equipment; or no one at all? What we need, concludes the Commissioner, are rules at EU level that establish data ownership.

5 The German automobile association ADAC has conducted tests showing that modern automobiles produce, process, store and forward vast amounts of machine-generated data; available at https://www.adac.de/infotestrat/-5mobiltechnik-und-zubehoer/fahrerassistenzsysteme/daten_im_auto/default.aspx.
controversial – database right. Whereas database right protects data on the double condition that the data are structured in a ‘database’ and the database is the result of ‘substantial investment’; the novel right would directly protect machine-generated data without any material prerequisite. As this article argues, introducing such an all-encompassing property right in data would seriously compromise the system of intellectual property law that currently exists in Europe. It would also contravene fundamental freedoms enshrined in the European Convention on Human Rights and the EU Charter, distort freedom of competition and freedom of services in the EU, restrict scientific freedoms and generally undercut the promise of big data for European economy and society. In sum, it would be a very bad idea.

This article starts (in Section 2) by briefly examining the background and stated aims of the proposed new right: why would there be a need for creating a property right in industrial data? And what would be its subject matter and scope? Section 3 looks at existing intellectual property regimes, inquires to what extent these extend to data, and speculates how a data property right in data might affect these regimes. Section 4 thereafter scrutinizes the data right from the broader perspective of fundamental rights and freedoms. Section 5 concludes.

Although creating a property right in data surely has additional ramifications outside these fields, in particular for the right of informational privacy (personal data protection), the focus of this article will be on the law of intellectual property. We will therefore not examine whether the law of data protection might already imply a property right in personal data.9 Nor shall we query whether the civil law concept of private property might be extended – or already extends – to (recorded) industrial data, and thus offer alternate protection to data sets.10 We shall also avoid discussing other doctrines in potential support of ‘data property’, such as criminal law12 or trade secret law13, and stay away from the contract and consumer law related issues of ‘trading’ personal data for services, which have become moot in the light of the proposed EU Digital Content Directive.14

Finally, a general caveat is in order. Whereas the European Commission has now posited the issue of ‘data property’ as worthy of serious discussion, the policy arguments advanced in favor of introducing such a right are underdeveloped, and its contours remain sketchy at best. Criticizing a right of data property is therefore taking aim at a moving target.

2. Quid data property?

The arguments advanced by proponents of introducing a right of ‘data property’ can be roughly summarized as follows: in the context of competition and innovation, especially in relation to the rise of ‘smart manufacturing’, which involves real-time exchanges of massive amounts of data between machines and robots; the economic potential of ‘mining’ Big Data (i.e. extracting in formation by way of sophisticated large-scale data analysis);15 and the promise of the Internet of Things, the magical world where machines quasi-independently communicate and exchange directly with other machines, such as the ‘intelligent’ energy meter that sends usage data to the energy company, or the axiomatic refrigerator that automatically orders milk and coffee from the online supermarket. As in Commissioner Oettinger’s op-ed, many of the examples used in the literature are taken from the automotive sector, where data have become essential input and valuable output in manufacturing and navigation.16 The specter of Google’s self-driving car potentially out-competing the European car industry is never far away.

Having thus demonstrated that data have tremendous and increasing value, proponents go on to point out that current legal regimes, such as traditional civil-law based property right and existing intellectual property regimes, do not, or do not adequately, protect these data. Admittedly, non-property regimes such as contracts and trade secret protection might occasionally be used in the literature are taken from the automotive sector, where data have become essential input and valuable output in manufacturing and navigation.16 The specter of Google’s self-driving car potentially out-competing the European car industry is never far away.

Finally, a general caveat is in order. Whereas the European Commission has now posited the issue of ‘data property’ as worthy of serious discussion, the policy arguments advanced in favor of introducing such a right are underdeveloped, and its contours remain sketchy at best. Criticizing a right of data property is therefore taking aim at a moving target.

23 Kerber (n. 21), 998. See also Drexl (n. 13).
ments for the introduction of such a new IPR.24 A more recent, and more elaborate study by the Joint Research Centre of the European Commission is somewhat less skeptical,25 recognizing that legal uncertainty regarding data ownership rights might negatively affect the efficiency of data markets. However, this study also concludes that there are, at present, no compelling economic arguments to advise regulatory intervention.

This article will, however, not further engage in economic analysis of a possible data property right, but focus instead on its consequences for the existing system of intellectual property. In order to do so, it is important to gain some preliminary understanding of what such a right might entail. Drawing from the sketch presented in the Staff Working Paper that underlies the European Commission’s recent Communication,26 which seems to be largely based on the work of Prof. Herbert Zech, we assume the features of a data producer’s right to be roughly as follows. The right would create a right in rem (i.e. a property right enforceable against the world) in respect of “non-personal or anonymised machine-generated data”. It would encompass “the exclusive right to utilise certain data, including the right to licence its usage. This would include a set of rights enforceable against any party independent of contractual relations thus preventing further use of data by third parties who have no right to use the data, including the right to claim damages for unjust enrichment and use of data without the right.”27

Whereas the Commission remains vague on the issue of initial ownership, according to Prof. Zech, the right would initially vest in “the economically responsible operator of equipment that generates the data (data producer)”.28 As the European Commission concedes, thus allocating the right might be highly problematic in practice, since data-generating machines are often owned and operated – and corresponding investments done – by numerous different actors.29

In view of its stated aims, the right would have to be fully transferable.30 And, from a strategic point of view, a proper distinction between data (raw data) and semantic data in the following section.

This article will, however, not further engage in economic analysis of a possible data property right, but focus instead on its consequences for the existing system of intellectual property. In order to do so, it is important to gain some preliminary understanding of what such a right might entail. Drawing from the sketch presented in the Staff Working Paper that underlies the European Commission’s recent Communication, which seems to be largely based on the work of Prof. Herbert Zech, we assume the features of a data producer’s right to be roughly as follows. The right would create a right in rem (i.e. a property right enforceable against the world) in respect of “non-personal or anonymised machine-generated data”. It would encompass “the exclusive right to utilise certain data, including the right to licence its usage. This would include a set of rights enforceable against any party independent of contractual relations thus preventing further use of data by third parties who have no right to use the data, including the right to claim damages for unjust enrichment and use of data without the right.”

Whereas the Commission remains vague on the issue of initial ownership, according to Prof. Zech, the right would initially vest in “the economically responsible operator of equipment that generates the data (data producer)”. As the European Commission concedes, thus allocating the right might be highly problematic in practice, since data-generating machines are often owned and operated – and corresponding investments done – by numerous different actors.

In view of its stated aims, the right would have to be fully transferable. And, from a strategic point of view, a proper distinction between data (raw data) and semantic data in the following section.
appears to extend copyright protection to machine-created works: “in the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken”.41 The European Court’s distinction with Football’ Database (‘fixture’ database) remains to be seen. According to the European Court clarified in British Horseracing,42 sui generis protection does not extend beyond misappropriation of data (contents) that result from substantial investment. In other words, the database right tolerates takings of (potentially valuable) data that are not the product of substantial investment.

Another delimiting factor is the notion of ‘database’. Art. 1(2) of the Directive defines this as “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”. While the Explanatory Memorandum generally describes the contents of the database as “information in the widest sense of that term”,43 the compiled data or materials must be ‘independent’, that is to say, “materials which are separable from one another without their informative, literary, artistic, musical or other value being affected”.44 Therefore an audiovisual, cinematographic, literary or musical work or a sound recording does not qualify as a database, even if it can be perceived as a representation of data (recital 17). This reflects a clear intention on the part of the European legislature to avoid extensive overlaps between the database right and existing copyright and neighbouring rights.45

Finally, according to art. 1(2) of the Database Directive, the individual elements of the database must be “arranged in a systematic or methodical way”. This squarely rules out protection – whether by copyright or by database right – of (collections of) raw machine-generated data.

3.3 Phonogram protection

In addition to copyright and database right, the phonographic right – one of the four neighbouring rights recognized at EU level – merits brief consideration. The rights of phonogram producers are harmonized by the Rental Right Directive (currently Directive 2006/115/EC) and the Information Society Directive (Directive 2001/29/EC). These Directives leave defining the notion of ‘phonogram’ to the WIPO Performances and Phonograms Treaty (WPPT) of 1996. According to the WPPT (art. 2) ‘phonogram’ means “the fixation of the sounds of a performance or of other sounds, or of a representation of sounds, other than in the form of a fixation incor

data.46 But where in this spectrum between purely synthetic data and data ‘observed’ should we place machine-generated data? The answer depends on the type of data that the machine processes. For example, sensor data produced by a radar system or observation satellite are likely therefore to qualify as ‘created’ data, while data produced by machines such as those which process seismic or aerial photographs are more likely to qualify as ‘observed’, data.47 According to the Court, “geographical information extracted from a topographic map by a third party so that that information may be used to produce and market another map retains, following its extraction, sufficient informative value to be classified as ‘independent materials’ of a ‘database’ within the meaning of that provision.” Freistaat Bayern v Verlag Esterbauer GmbH, CJEU 29 October 2015, Case C-490/14.

42 Football Dataco and others (n. 39).
43 The Newspaper Licensing Agency & others v Meltwater & the PRCA, High Court of Justice (Chancery Division), 26 November 2010, [2010] EWHC 3099.
44 Bundesgerichtshof (‘Federal Supreme Court’), 1 December 2010, case I ZR 196/08.
46 British Horseracing (n. 45), para. 31.
porated in a cinematographic or other audiovisual work". By including ‘other sounds’ and ‘a representation of sounds’ this definition apparently encompasses raw audio data stored (‘fixed’) on a digital medium.

Whether there is a threshold criterion for the phonographic right that might delimit both the substance and scope of the right, is as yet unsettled under EU law. In its 2008 Metall auf Metall decision the German Constitutional Court extended neighboring rights protection to every single recorded note of a sound recording, because the record producer’s investment is reflected in every – even very minor – part of the recording.54 This suggests that no threshold criterion (no investment minimum) would apply. In a follow-up decision the German Constitutional Court has however held that a phonographic right of unlimited scope, as contemplated by the Federal Court, may collide with the ‘freedom of art’ that is constitutionally guaranteed in German law.55 Most recently, the Bundesarbeitsgericht has referred questions regarding the scope and limitations of the phonographic right to the EU Court of Justice.56

3.4 Assessment: impact of data property on the system of intellectual property law

As this section shows, both copyright and database right do not extend to data per se. Both regimes do conditionally offer protection to data compilations that result from creative selection and arrangement (copyright) or substantive investment (database right). Both regimes deny protection to raw data. For copyright, this follows from the axiom that only acts of authorship conducted by human beings are protectable. For database right, this is a consequence of the sui generis right’s categorial delimitation: only data structured in a ‘database’ qualify for protection. Moreover, the sui generis right’s substantial investment test sets an – admittedly fairly low – minimum threshold. If operating a machine that records sensor data does not amount (for example, a digital weather station or a bicycle computer), then this will not result in a protected database. The CJEU’s ‘Fixtures’ decisions pose an additional hurdle to sui generis protection for machine-generated data by excluding ‘created’ data from protection, thus ruling out machine-generated synthetic data.

In sum, introducing a right in raw, machine-generated industrial data, as envisaged in the Commission’s Communication, would go far beyond the main intellectual property regimes presently existing in Europe in the field of data and information, copyright and database right. Disruptive overlaps

How would this affect existing intellectual property law? In the first place, creating a new layer of rights in machine-generated data would cause broad and disruptive overlaps with copyright and sui generis right in productions made with the aid of digital machines. For example, a film shot with a digital camera would qualify not only as a work protected by copyright, but also as machine-generated (sensor) data subject to a ‘data producer’s right’. Similarly, the aggregate stock market database in a financial database would be protected both by sui generis right and a ‘data producer’s right’, since the data are recorded automatically by the computerized stock exchange.

Whereas the EU legislature has clearly intended to prevent the database right from spilling over into the realms of copyright and neighbouring rights, the ‘data producer’s right’ would lead to extensive overlaps. As a consequence, the new right might give rise to multiple competing claims of ownership in the same content. To continue with our first example, while the creators of the film (e.g., the director, screen writer, and other creators of the film) could claim authorship in the cinematographic work, the owner or operator of the computer might claim ‘data property’ in the photographic data (i.e., the digital representation of the film), – surely, to the unpleasant surprise of the film’s producer. Similar examples might be given with regard to digital photographs or e-books. In the second example, the database producer might be confronted with ‘data property’ claims of the stock exchange, or the exchange’s computational services company. Another consequence of this wide-ranging overlap would be that statutory limitations and exceptions under copyright, neighbouring rights or database right are ‘trumped’ by data producer’s right. For example, both copyright and database right in the EU presently allow users to copy or extract data from databases for non-commercial research purposes. Unless, the ‘data producer’s right’ would replicate all relevant existing exceptions, it would underscore these essential user freedoms.

This is especially true for data mining. Strangely, while the Commission’s Communication on ‘Building a European data economy’ ponders the introduction of an exclusive right in machine-generated data, one of the highlights of the DSM Directive proposal that is currently being debated in the European Parliament is a mandatory exception, both under copyright and database right, for text and data mining by non-commercial research organisations.57 In line with Prof. Zech’s suggestions, the European Commission in its Staff Working Paper attempts to alleviate concerns of wholesale overlap by distinguishing semantic from syntactic data. The proposed ‘data producer’s right’ would be conceived in such a way that “only the syntactical level of information is protected, not the semantic level”!58 What is probably meant here is that the raw data would be protected only as regards its digital representation (the machine-readable bits and bytes, the ‘ones and zeros’ in the digital file), not the informational content that these data convey. Thus, the European Commission hopes, the new right would not extend to ideas and information, and the new right would not become a ‘super-IP right’.59 But would such a distinctly real prevent the new right from extensively overlapping with existing IP rights? I do not believe so. The problem here is that digital data are commonly coded and interpreted following standardized rules and protocols. In other words, there usually will be a one-on-one relationship between the (syntactic) data substrate and the (semantic) content layer. Returning to our example of the digitally produced film, any copy of the film’s digital file (the syntactic data) would by necessity also reproduce the copyright protected work (the semantic layer). Thus, the new data right could be invoked against any digital copying (or streaming) of the digitized copyright work. For the same reason, the new right would broadly overlap with database right, even if its scope were confined to the syntactic layer. The phonographic right discussed above illustrates this point. Whereas its subject matter, like the proposed ‘data producer’s right’, is limited to the recorded signal (i.e., syntactic audio data), its scope extends into the semantic realm. Reproducing a cd recording of a musical performance will, by necessity, reproduce the underlying musical work and performance.

The only way to prevent the data right from becoming an all-encompassing ‘super-IP right’ would be to categorically exclude all data that (possibly) represent subject matter protected under traditional IP regimes: not just copyright, database right and neighbouring rights, but also design right and perhaps even patents. But even a non-overlapping data right would have seriously corrosive effects on the system of intellectual property, for various reasons. First, it would undermine the economic incentives that underlie IP rights. For example, the main rationale of the data protection is to promote the building and development of new and existing data and other materials. This incentive is clearly undercut if a lower-tier, no-threshold right in machine-generated data were to exist in parallel. Second, and more importantly, it would compromise the general principle of intellectual property – whether utilitarian or grounded in natural law theory – that protection be reserved to creation, innovation or otherwise meritorious investment. A data right in all data produced by machines might, on occasion, protect assets of considerably economic value, but nothing of merit. This has ramifications, in particular, at the political level. With intellectual property laws under increasing fire, legislatures – at EU and national level – need powerful and convincing arguments to defend existing regimes and introduce new rights. In this volatile political climate proposing a data producer’s right with the

54 Metall auf Metall, Federal Supreme Court, 20 November 2008, case I ZR 112/06, GRUR 2009, 403; German Constitutional Court, 31 October 2017, case 1 BvR 402/15.
55 Metall auf Metall III, Federal Supreme Court, 1 June 2017, case I ZR 115/16.
56 European Commission, Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market, Brussels, 14 September 2016, COM(2016) 593 final. Art. 3(1) of the proposed Directive provides: “Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(6) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research.”
57 Zech, ‘Informations as a tradable commodity’ (n. 19).
58 European Commission, Staff Working Document (n. 4), 35.
59 European Commission, Staff Working Document (n. 4), 34.
sole aim of (better) protecting the economic assets of the automotive (or any other) industry will surely backfire. Not only is such an initiative likely to fail in the legislative process, but it will also (re)ignite broader discussions on the legitimacy of intellectual property law.

No legal certainty

Another, more mundane objection against a property right in data lies in its inherent lack of legal certainty. Although it is still not fully conceptualized, it is difficult to imagine a data right sufficiently stable in terms of subject matter, scope and ownership to be admitted to the ranks of intellectual property. As to subject matter, if the right vests in data generated by machine processes, which data would it protect? All the data that the machine produces within a given time frame (e.g. an hour, a minute or a second)? Or all the data that result from a finite machine process (e.g. all the data gathered by a satellite that sensors the earth)?

Admittedly, the sui generis database right has already raised similar questions. With data in a database constantly being updated, what exactly constitutes the protected database? But in database law the definition of ‘database’ and requirement of substantial investment create at least some measure of permanency in the subject matter and scope of the right. This stability is, however, completely absent from the data producer’s right. The problem here is that industrial data generation mostly occurs in real time. The ‘velocity’ – the dynamic nature – of big data makes it almost impossible to identify and protect the intellectual property.

A related problem is allocating ownership of the right. Since the right would be sparked by the data generation process – and, by implication, its scope of protection – with sufficient legal certainty.

The exclusion of data per se from the scope of existing intellectual property regimes is not a case of ‘copyright in information’ (1989).

As to subject matter, if the right vests in data generated by machine processes, which data would it protect? All the data that the machine produces within a given time frame (e.g. an hour, a minute or a second)? Or all the data that result from a finite machine process (e.g. all the data gathered by a satellite that sensors the earth)?

Admittedly, the sui generis database right has already raised similar questions. With data in a database constantly being updated, what exactly constitutes the protected database? But in database law the definition of ‘database’ and requirement of substantial investment create at least some measure of permanency in the subject matter and scope of the right. This stability is, however, completely absent from the data producer’s right. The problem here is that industrial data generation mostly occurs in real time. The ‘velocity’ – the dynamic nature – of big data makes it almost impossible to identify and protect the intellectual property.

A related problem is allocating ownership of the right. Since the right would be sparked by the data generation process – and, by implication, its scope of protection – with sufficient legal certainty.

The exclusion of data per se from the scope of existing intellectual property regimes is not a case of ‘copyright in information’ (1989).

4. Data property and the free flow of information

The exclusion of data per se from the scope of existing intellectual property regimes is not merely ontological. Although old-school author’s right scholars might argue that data are not copyright works, because data are not ‘created’, this is at best a partial explanation for this exclusion. Rather, IP law’s abhorrence of protecting data reflects implicit or explicit information policies not to protect data. These policies are, in turn, informed by a variety of public interest values and concerns. In the first place, of course, freedom of expression and information – the fundamental freedom enshrined in the European Convention on Human Rights (art. 10 ECHR) and the EU Charter (art. 11).

As case law and doctrine regarding the Convention teach, this fundamental freedom is to be interpreted broadly. Article 10 ECHR is phrased in media-neutral terms and thus applies to old and new media alike. The term ‘information’ (in French: ‘informations’) encompasses, at the very least, the communication of facts, news, knowledge and scientific information. It also, undoubtedly, extends to syntactic data; the scope of article 10 is not limited to (semantic) speech, but extends to the means used for communication purposes. To what extent the article’s protection extends to commercial speech has been a matter of some controversy. However, the European Court of Human Rights has made it clear that information of a commercial nature is indeed protected, albeit to a lesser degree than political speech.62

Article 10 ECHR prevents states from creating restrictions to the free flow of information unless such restrictions “are prescribed by law and are necessary in a democratic society [...] for the protection of the [...] rights of others”. From this perspective data and information must flow freely, uninhibited by property rights or other state-created restrictions, unless a compelling societal need for protection (“necessary in a democratic society”) can be established. Freedom of expression and information, in other words, makes intellectual property rights in data the exception to the default rule of freedom.63

This brings us back to the question of expediency. The EU legislature would bear the burden of proving that a property right in machine-generated data is a socially and economically justified ‘fair’ interference in the (informations) freedom of European citizens and companies. It would have to convincingly support the need to access and reutilize machine-generated data. In light of the abundant praise in political literature of ‘big data’64 and big data mining as drivers of progress and prosperity, and the absence of convincing evidence supporting a property right in machine-generated data, this burden of proof would well be too high. Furthermore, the free flow of information militates strongly against any new right of intellectual property that would restrict scientists’ access to data - a freedom that the EU legislature expressly wishes to preserve as regards ‘text and data mining’ by non-commercial research institutions. Note that this freedom finds additional support in art. 13 of the EU Charter ("The arts and scientific research shall be free of constraint. Academic freedom shall be respected"). Another area where a data right would patently conflict with freedom of expression and information is journalism, where mining data has become an essential tool for investigative reporting.65

A second over-arching policy consideration underlying intellectual property law’s reluctance to protect data per se is freedom of competition (enshrined in art. 16 of the EU Charter as the “right to conduct a business”). The economic potential of ‘big data’ demonstrates, machine-generated data are both input and output to innovative manufacturing processes and value-added services, and thus a major driver of economic growth. This calls for measures promoting access to data and freedom to use data rather than commodification of data by creating property rights in data. Unless equipped with wide-ranging exceptions and safety valves, introducing a new property right in data might create undesirable data monopolies that could impede, rather than foster, competition in this rapidly evolving European ‘data market’ place.66 At the global level, introducing data property rights in the EU might well lead to anti-competitive distortions as well, in cases where European data users are obliged to purchase licenses for usage of data freely available to their competitors in the United States.

63 See (for copyright) Ashby Donald and Others v France, European Court of Human Rights 10 January 2013, No. 36769/08; ECLI:2013:0101JUDU0367690.
65 See Dammann v. Switzerland, ECHR 25 April 2006, no. 77551/01. The Court opines that “the gathering of information was an essential preparatory step in journalism and an inherent, protected part of press freedom”. Surprisingly, data mining for journalistic purposes seems to be overlooked in the proposed TDM exception of the DSM Directive. 66 Max Planck Institute for Innovation and Competition, ‘Data Ownership and Access to Data’, Position Statement of the Max Planck Institute for Innovation and Competition, Max Planck Institute for Innovation & Competition Research Paper No. 16-10 (2016). 2.
Finally, a novel data right would also create new barriers to the freedom of services, one of the four freedoms of the EU Internal Market. In its Communication on ‘Building A European Data Economy’, the European Commission interprets this freedom, together with the freedom of establishment, as implying a “principle of free movement of data within the ELP”. It is hard to see how a novel property right in machine-generated data would square with this freedom.

5. Conclusion

This article makes the case against introducing a data property right. As we have seen, there are abundant reasons to reject this idea. A ‘data producer’s right’ in machine-generated data would ride roughshod over the existing system of intellectual property. It would violate one of the IP system’s main maxims that data per se are “free as the air for common use”, and that only creative, innovative or other meritorious investment is protected. It would corrode IP’s mechanism of incentives by creating an underlayer of rights that automatically protects all data produced with the aid of machines. This parallel layer of rights would, most likely, extensively overlap with other IP regimes, and thus create undue impediments for the exploitation of existing rights, such as copyright and database right, and endanger user freedoms guaranteed under these regimes. As this article has shown, the ‘velocity’ of real-time data generation makes it difficult, or even impossible, to circumscribe its subject matter, scope of protection and ownership. More generally, a property right in machine-generated data would contravene freedom of expression and information, and pose new obstacles to freedom of competition, freedom of services and the ‘free flow of data’.

The great promise of big data – for the economy, for science, for society at large – is that this resource may be freely exploited. Introducing a ‘data right’ preventing unauthorized access to big data would directly contradict this. Indeed, it is hard to understand how the proposed new right would square with the text and data mining proposed by the European legislature in the current EU copyright reform package.

If, as the European Commission rightly believes, “big data, cloud services and the Internet of Things are central to the EU’s competitiveness”, one would have expected supporters of a novel data producer’s right to present powerful and convincing arguments in support of this revolutionary proposition. So far, the case for a property right in machine-generated data has yet to be made. As Prof. Drexl and others have pointed out, the existing toolkit of trade secret protection, contract and technological protection measures offers data producers ample means of securing de jure or de facto exclusivity. Rather than wasting time and effort on inventing a data producer’s right, the focus of the European Commission’s possible interventions should be on fostering access to big data.

Fortunately, the possible introduction of a ‘data producer’s right’ is only one of several policy options currently being contemplated by the Commission in its ‘European Data Economy’ initiative. As this article has shown, there are innumerable reasons for the European Commission not to go down this road. If nothing else, Europe’s experience with the sui generis database right should give reason for extreme caution. In 2005, less than ten years after it was introduced at EU level, the European Commission published its first review of the Database Directive, a remarkably self-critical assessment. According to the Commission, “[t]he economic impact of the ‘sui generis’ right on database production is unproven. Introduced to stimulate the production of databases in Europe, the new instrument has had no proven impact on the production of databases”. The Commission’s report also suggests that the sui generis right has not helped the European industry to overcome its productivity gap vis-à-vis the United States. It points to several other deficiencies of the sui generis right, such as its uncertain contours, and its proximity to a property right in data that might negatively affect innovation and growth. The report juxtaposes the legal situation in the EU with that in the United States, where since the Supreme Court’s landmark Feist decision, no legal protection for ‘sweat of the brow’ based databases exists. Nevertheless, as the Commission wryly observes, “there has been a considerable growth in database production in the US, whereas, in the EU, the introduction of ‘sui generis’ protection appears to have had the opposite effect.”

The 2005 evaluation report concludes by offering four possible ways forward: (1) repeal the whole Directive; (2) withdraw the sui generis right, (3) amend the sui generis to clarify its scope, and (4) maintain the status quo. Despite these harsh conclusions, the database right has yet to be amended or repealed. The problem is that removing (parts of) a directive is, politically and legislatively, even more complex than substantive harmonization. Repealing the database right would require a new directive not only rescinding major parts of the existing Directive, but also – absurdly – instructing Member States to abolish sui generis database protection. Unsurprisingly, the only option that has so far materialized from the Commission’s assessment is no. 4: “do nothing”.

The lessons of the EU’s database experiment are not to be forgotten. Introducing a novel right of intellectual property should never be done in the spur of the moment. Any new right should be contemplated only after conducting thorough economic, evidence-based research that demonstrates a real need for the right and predicts its consequences for information markets and society at large. Assuming a convincing case in support of the right might indeed be made, this should then be followed by systematic legal analysis of the new right’s contours and scope, and of its impact on the existing system of intellectual property. The two-tiered structure of the Union does not allow for legal experimentation at the EU level. Like the database right, a ‘data producer’s right’ would be here to stay – a most unwelcome guest in the house of European intellectual property.

67 European Commission, ‘Building A European Data Economy’ (n. 4), 7.
68 DSM Strategy, p. 34.
69 Drexl (n. 13), 66.
70 See Drexl (n. 13), 41ff; Max Planck Institute Position Statement (n. 66); see also European Commission, Staff Working Document (n. 4), 36ff.
REMUNERATION OF CONTENT CREATORS IN THE DIGITAL SPACE:

CHALLENGES, OBSTACLES AND A COMMON LANGUAGE TO FOSTER ECONOMIC SUSTAINABILITY AND CULTURAL DIVERSITY

Dr Giuseppe Mazziotti
Discussion paper, 7-8 February 2019
Acknowledgements

This document has been prepared for the Department of Canadian Heritage and the Canadian Commission for UNESCO by Giuseppe Mazziotti. The views, opinions and recommendations expressed in this report are those of the author and do not necessarily reflect the official policy or position of the Government of Canada. Responsibility for any errors, interpretations or omissions lies solely with the author.
# TABLE OF CONTENTS

1. **Introduction** ............................................................................................................... 4
2. **Context: online content dissemination and the place of creators** ................................. 5
   2.1. Unauthorised file-sharing and peer-to-peer software ............................................. 6
   2.2. Content-sharing services: social media and user-generated content platforms ....... 6
   2.3. On-demand content platforms ............................................................................. 7
3. **Observed economic challenges** .................................................................................. 8
   3.1. Evolution of online piracy ..................................................................................... 8
   3.2. Content-sharing platforms .................................................................................. 9
   3.3. On-demand content services .............................................................................. 11
4. **What has been done so far? Responses from governments** ......................................... 12
   4.1. International treaties and their broadly protective approach to copyright ............ 12
   4.2. Copyright’s territoriality ....................................................................................... 12
   4.3. Copyright and freedom of contract ..................................................................... 13
   4.4. Empowering creators through transparency of information .................................. 15
   4.5. Collective management of authors’ rights in the music sector ............................... 15
   4.6. Online copyright enforcement ............................................................................ 17
5. **Responses from online platforms and civil society** .................................................... 19
   5.1. Online platforms ................................................................................................. 19
   5.2. Civil society ........................................................................................................ 20
6. **Going forward: major obstacles** .................................................................................... 21
   6.1. Secrecy, lack of transparency and scant data on content value chains .................... 21
   6.2. Absence of rights management information standards ......................................... 22
   6.3. Bargaining power, size and origin of the largest online platforms .......................... 22
   6.4. Risks and costs entailed by online copyright enforcement ..................................... 23
7. **“Remuneration” and content “creators”**: towards a shared understanding and a common language at international level ......................................................................................... 24
   7.1. Notion of “creation” and different kinds of contributors to content value chains ...... 24
   7.2. Function (and limits) of copyright law .................................................................. 24
   7.3. Suggested elements of a multilateral dialogue ....................................................... 25

**Bibliography** .................................................................................................................... 29
Executive Summary

The Internet has subverted the traditional business models of content industries and has allowed online intermediaries to dominate the dissemination and commercial exploitation of knowledge, culture and entertainment. In supporting the creation and dissemination of French-language cultural content, the Joint Declaration on Cultural Diversity and the Digital Space, signed by the Canadian and French governments in April 2018, raises a number of questions on the future of technology, cultural and copyright policies. In particular: What is the role of copyright and remuneration in preserving and promoting diverse cultural creation at a time when digital markets exacerbate pre-existing “winner-takes-all” dynamics with respect to success and distribution of different types of creative works? What is the place of creators and what is the role of digital platforms in online dissemination of cultural expressions? What conception of content “creators” will ensure economic sustainability and diversity of cultural creation in the long term? This paper seeks to answer these questions.

Web-based communications have rapidly changed the context and the conditions under which creative works have been produced and disseminated. In the Internet’s infancy, peer-to-peer dissemination of music files through file-sharing networks raised the issue of how intellectual works could be remunerated in the absence of new models for non-physical transactions. At that time, one idea was that cultural content would be disseminated in a disintermediated way, without the possibility of remunerating content creators for online exploitations of their works. Very few services, starting with Apple’s iTunes in 2001, offered copyright content legitimately. The emergence of content-sharing services, such as social networks and user-generated content platforms, as well as download and streaming services, gave rise to a process of re-intermediation in content distribution. The quick rise of digital platforms that make available works uploaded by their users inevitably raised policy issues concerning online intermediary liability for copyright infringement. All of the legal systems this paper takes into consideration — Canadian, US and EU — protect online platforms’ neutrality by exempting platforms from copyright liability insofar as they remove (or record) infringements in response to copyright holders’ notifications (so-called “notice-and-takedown” or, in Canada, “notice-and-notice” mechanisms).

An environment where online piracy remains rampant (although its impact on sales is uncertain) and online platforms give access to copyright works either for free (via content-sharing platforms) or through subscription fees to access vast collections (via streaming services) inevitably triggers economic challenges to remuneration of content creators. For creators, these challenges are essentially those of not being remunerated at all or being compensated very little, because of the uncertain or very low commercial value of the vast majority of creative works on digital platforms. Even though the analysis relies on a broad notion of content “creators,” which encompasses all copyright holders and the creative sector as a whole, the paper takes into special consideration individual authors and performers and their position vis-à-vis content-sharing platforms and content producers. The economic situation of individual creators is particularly important for the purpose of this paper since diversity of cultural expressions depends essentially on artistic and intellectual labour of individuals (or groups of individuals) rather than on investments and businesses of enterprises and cultural industries. With the significant exception of Google and its “Content ID” technology used across the YouTube platform, most social media platforms do not facilitate copyright enforcement and do not give individuals the possibility of monetizing online exploitations of their works. Notice-and-takedown systems work much better for wealthier rights-holders than for individuals or small-size content producers, who do not have time or resources to monitor what Internet users upload on social media. Moreover, licensed content platforms give rise to scalable and very unequal environments where a very few superstars have a disproportionately high share of the market (and even the stars earn next to nothing per stream).

Governments have attempted to tackle these issues and ensure some fairness in remuneration of creators, as well as
transparency in the ways creative works are exploited. Regulatory interventions have targeted crucial aspects such as the intersection of copyright and contract law and the establishment of limitations to authors’ and performers’ contractual freedom to sell all of their rights to publishers and other content producers without benefiting effectively from the revenues generated by their works. To this end, in jurisdictions such as the EU member states individual creators are increasingly placed in a position to exercise, under different conditions, rights to termination of their copyright transfers and rights to obtain information on the different exploitations of their works and the related revenues. In Europe, for instance, authors and performers can also rely on an improved functioning of collecting societies for the licensing of digital uses and on freedom to choose a rights manager of their choice, independently of their place of residence or country of origin.

Governments have also sought to improve the effectiveness of copyright enforcement through multilateral and national initiatives targeted at both structurally infringing websites (such as sites implementing sophisticated peer-to-peer technologies such as The Pirate Bay) and the largest online platforms. In this regard, the EU and the US seem to have different approaches to the problem of remuneration of creative works exploited across online platforms. The US still relies on its liability exemptions (or “safe harbours”) to give digital platforms such as YouTube, Facebook and Twitter the shield of immunity following notice-and-takedown procedures. The EU, instead, is currently reconsidering the principle of platforms’ neutrality and seeking to oblige the platforms — in light of their active role in optimising presentation of the uploaded works or promoting them — to obtain a license and pay for the contents their users upload. Platforms and civil society organizations have not remained inactive on the copyright enforcement front. YouTube, Facebook and other platforms have become increasingly compliant with copyright through content identification technologies and rights management software that allow them to filter unauthorised works and let content creators decide whether the works should be removed from the platforms or monetised. Civil society organizations such as Creative Commons have contributed to the development and adoption of technologies and licensing standards that help content creators, online intermediaries and Internet users understand whether a copyright work is made available to the public for profit or for free, for the purpose to be shared with others.

The complex scenarios the paper describes reveal the existence of major obstacles for better remuneration of content creators: (i) secrecy and lack of data on how the largest online platforms extract value from content-related interactions with their users and from imposition of unfair conditions to content creators; (ii) absence of standards of rights management information in each creative sector, which would facilitate licences with (and payments from) digital content exploiters; (iii) the bargaining power and size of the largest online platforms, that has been described as a threat to democracy and a natural target of antitrust enforcement; and (iv) the risks and social costs triggered by online enforcement measures with regard to freedom of expression and communication, net neutrality and freedom to do business online. A multi-stakeholder dialogue in the context of an international policy initiative could help develop a common language on remuneration, long-term sustainability of content creation and cultural diversity. What is indispensable is a reconciliation of the copyright aspects related to trade with other cultural and media policies that are expressly contemplated in the 2005 UNESCO Convention on cultural diversity. This reconciliation is essential if copyright is to regain its centrality and credibility in the Internet-related political debates.

1. Introduction

In April of 2018 the Canadian and French governments signed the Joint Declaration on Cultural Diversity in the Digital Space, in accordance with the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions. The declaration recalls that cultural diversity is inseparable from human rights and fundamental freedoms such as freedom of expression, communication and the possibility for individuals to choose their cultural and linguistic expressions. In emphasizing their common will to support creation and dissemination of French-language cultural content, the two gov-
ernments agree that States, digital platforms and civil society should contribute to the economic sustainability of content creators and to respect for copyright. It has become evident that the Internet has subverted the traditional business models of content industries and has allowed online intermediaries such as digital platforms to dominate the market. From an economic point of view, digital markets and online platforms exacerbate pre-existing characteristics of inequality of success and distribution of works and disparities of income among different authors, works and repertoires. While supporting Internet neutrality as well as sustainability of content creation through fair remuneration and copyright enforcement, the joint declaration raises a number of questions: What is the place of creators in today’s digital world? What concept of content “creators” should policy makers take into consideration to ensure economic sustainability of cultural creation and diversity of cultural expressions? What is the role of remuneration in preserving and promoting creation at a time when access to knowledge, culture and entertainment occurs increasingly online? And what is the role of copyright?

This paper seeks to address these questions. Section 2 describes the context in which creative works have been disseminated following the advent of web-based communications: from disintermediated forms of content sharing enabled by peer-to-peer software to the emergence of social media, user-generated content platforms and streaming services. Section 3 identifies emerging economic challenges for remuneration of content creators at a time when online piracy remains very relevant and web-based platforms give access to repertoires and vast collections of works for free (via content-sharing platforms) or on the grounds of a subscription and payment of a monthly fee (via download and streaming services). Section 4 considers how governments have addressed the issue of copyright protection and exercise of a plurality of rights that, even in the digital space, should ensure remuneration of cultural creation and content distribution. Regulatory interventions have targeted aspects of these issues such as individual and collective management of copyright, contractual agreements and transfers of rights from authors and performers to content producers and other attempts to ensure transparency and fairness across value chains of content production. Section 5 shows how online platforms and civil society have responded to the radical changes digital creation and online content dissemination have entailed by adopting technologies and licensing standards that help content creators, intermediaries and Internet users understand the conditions under which works are made available to the public. Section 6 focuses on major obstacles towards the achievement of conditions under which remuneration of creative works might regain centrality: for instance, secrecy and lack of data on how dominant (if not monopolistic) online platforms extract value from content-related interactions with their users and the imposition of unfair conditions to content creators; and, as a result, a low or very uncertain commercial value of average digital works. Finally, Section 7 provides a reflection and policy suggestions on whether and how a shared understanding and a common language on remuneration and long-term sustainability of content creation can be developed at international level. In particular, this section considers (i) whether a reconciliation of the aspects of copyright related to trade and culture is possible and (ii) whether a multi-stakeholder dialogue can help develop an interface between international copyright agreements and instruments such as the 2005 UNESCO Convention on the protection and promotion of the diversity of cultural expressions.

2. Context: online content dissemination and the place of creators

Digitisation of information and the advent of the Internet — as an unprecedented, borderless and decentralised medium of expression and communication — revolutionised the way individuals and cultural industries produce and disseminate ideas and creative works. In the mid-1990s, digitisation of information exchanged on the Internet and the end-to-end design of this new medium triggered a debate on whether or not copyright could survive. Some writers predicted that, in the absence of successful new models for non-physical transactions, there would have been no way to assure reli-
able payment for intellectual works (Barlow, 1994). Other scholars, to the contrary, were convinced that the new digital environment would give authors greater opportunities to trace consumption of their works and to gain remuneration through micropayments, as if the Internet could become a “celestial jukebox” (Goldstein, 1994).

2.1. Unauthorised file-sharing and peer-to-peer software

In the Internet’s infancy, due to limited bandwidth and slow data processing, this new medium did not allow transmissions of large amounts of information. However, new technologies such as audio compression formats and peer-to-peer software started allowing Internet users to share sound recordings with each other for free, bypassing intermediation of record producers, skipping payment of remuneration and challenging the enforcement of copyright. For a number of years file-sharing threatened the survival of the recording industry since music files shared for free had the potential to replace CDs and other physical formats, which were the core business of that industry. To escape liability and ensure better performances, new platforms and file-sharing protocols relying on sophisticated technologies (for instance Napster, Grokster, eMule and BitTorrent) facilitated direct exchanges between users without storing copyright-protected works on their servers. As soon as bandwidth enabled faster and larger content transmissions, such technologies and protocols started targeting films, TV series and video games (Quintais, 2018).

The practice of sharing copyright works without intermediation rose to such prominence in the Napster and Grokster era that influential academics, in slightly different ways, proposed legalization of file-sharing. Their main idea was that permitting non-commercial sharing of online works by requiring payments to content creators via Internet access providers would ensure remuneration for creators without hindering web-based communication (Netanel, 2003; Fisher, 2004). To measure user demand and ensure remuneration proportionate to effective use of these works, their solutions presupposed either registration of the works with a government agency (and a subsequent incorporation of fingerprints into the content files) and/or periodic surveys and inquiries aimed at metering uses of registered works. The strongest objection to this idea was that such a broad statutory licensing scheme would discourage formation of new markets and the emergence of innovative services based on property rights and customised licenses (Merges 2004).

2.2. Content-sharing services: social media and user-generated content platforms

Until the launch of Apple’s iTunes music store in 2001, file-sharing was the most popular way to access copyright works on the Internet. Following, the rise of on-demand content stores and streaming services, together with the emergence and large-scale diffusion of social networks and Web 2.0 technologies, triggered a process of re-intermediation in digital content distribution (Renda et al., 2015).

2.2.1. Re-intermediation in content distribution

The rise to prominence of video-sharing platforms such as YouTube and Vimeo, social networks like Facebook and Twitter and other interactive services or dedicated platforms for photos (for instance Instagram, Flickr and Pinterest) and sound recordings (for example Soundcloud) has significantly expanded the opportunities for Internet users to access creative works. An essential feature of these platforms, from the perspective of cultural creation and remuneration, is that they are not designed to allow or to facilitate a distinction between original creations of the platform user and works created by someone else which are uploaded by the user without the right-holder’s authorisation. From a legal perspective, access to and use of such platforms is conditional upon the acceptance of terms and conditions that oblige subscribers not to share and publish works created by third parties without their authorisation. However, from the outset, providers of content-sharing platforms have been reluctant to enforce this contractual condition and to monitor the contents their subscribers upload.
2.2.2. Protection of online platforms’ neutrality

It would be impossible to understand the conduct and policies of content-sharing service providers without considering the special treatment and immunity that laws such as the 1998 Digital Millennium Copyright Act in the US, the 2000 e-Commerce Directive in the European Union and, at a later stage, the 2012 Copyright Modernization Act in Canada granted to Internet service providers, in particular to suppliers of “hosting” services. US and EU laws created, from the late 1990s onwards, liability exemptions (or “safe harbours”) that made hosting service providers not liable for activities carried out by their users if, after gaining knowledge of unlawful conduct, the service providers promptly removed illegal materials. Before adopting ad hoc legislation on online intermediary liability for copyright infringement in 2012, Canada had achieved an equivalent result through a 2000 agreement by the Canadian Association of Internet Providers (CAIP) and the music and cable industry, which successfully dealt with copyright infringement claims. The agreement sought to put into practice a solution enshrined in a judgment of the Canadian Supreme Court holding that Internet service providers might have incurred secondary liability if they had notice of a potential copyright infringement carried out by its customers and did not take remedial action. The 2012 legislation is basically a codification of the rules suggested by the Supreme Court.

The main idea, justified also by the need to defend the neutral design of the Internet, was that Internet service providers should not be expected to monitor the traffic end-users delivered or received through their networks. Content sharing services have taken advantage of these exemptions to skip liability and to remove unauthorised copyright works uploaded by their subscribers only after having been informed and requested to do so (through a ‘notice’). The consequence of this exemption, in Europe and in the US, has been a broad implementation of so-called “notice-and-takedown” mechanisms to social media and user-generated content platforms. In Canada, instead, the liability exemption led to the application of a lighter “notice and notice” regime, with no removal of the infringing materials: copyright holders send a detailed notice to report and locate an infringement of their copyright and then the service provider forwards it to the accused subscriber, keeping a record of it.

Copyright enforcement across content-sharing platforms became largely dependent on whether or not rights-holders had the resources and the possibility of monitoring uploads and requesting the takedown of their works. More recently, enforcement has started depending on whether a given platform implements forms of content identification to remove copyright-infringing materials and other harmful content (such as hate speech). Nonetheless, it is evident that, at least at the beginning, this trend inevitably transformed content-sharing platforms into de facto media companies whose unrestricted communication of large amounts of unlicensed copyright works to the public can be stopped or monetized only via ex post initiatives of rights-holders.

2.3. On-demand content platforms

From the perspective of content creators, on-demand download and streaming services are definitely a better option than content-sharing platforms since all works are licensed ex ante and their use is remunerated through fees that content producers and/or authors’ collecting societies negotiate with each service provider. On-demand content platforms act as intermediaries between traditional creative industries and consumers. Apple’s iTunes was the first service of this kind. These services have very different features and business models. Some of the biggest content platforms function as large retailers who sell permanent digital copies (downloads) of copyright works. These platforms often combine online catalogues with the sale of dedicated devices (for instance, Amazon’s Kindle or Apple’s portable hardware) enabling consumers to access and enjoy the works they buy. Other platforms, such as Spotify, Netflix and Amazon, work as subscription-based radio and television services, giving access to music repertoires or collection of films, TV programs and other audiovisual works.

One of the distinctive features of online platforms is their ability to know and exploit user or consumer preferences and attention. Through their websites and interfaces, these service providers collect and store personal data whenever a consumer buys a product or a subscriber uses one of their
service features. Such an extensive knowledge of their subscribers’ preferences places online platforms not only in a position to sell and earn revenues from online advertisers, as occurs in the case of content sharing services, but also to target commercial offerings at the single consumer. User profiling allows on-demand content suppliers to take advantage of known consumer preferences in a way that reflects the consumer’s behaviour on the platform.

3. Observed economic challenges

The digital space has raised challenges to remuneration of content creators in multiple ways. This section identifies the main economic challenges triggered by different distribution models developed in the last two decades. Going from the worst to best case scenarios for content creators, this section considers the evolution of file-sharing and the still significant rates of online piracy (§3.1); the making available of copyright works on content-sharing platforms (§3.2); and online distribution via on-demand content services (§3.3).

3.1. Evolution of online piracy

Since the infancy of web-based communications, disintermediated forms of online content communication, beginning with the file-sharing of sound recordings in the late 1990s, have raised unprecedented threats to remuneration of content creators. Due to the Internet’s architecture and the technology that enables massive copying of digital files without degradation in quality, copyright owners lost control over large-scale use and exponential distribution of their works (Renda, 2011).

3.1.1. Piracy and growth of legal and illegal streaming services

Available studies and data (Kantar Media, 2016) on user conduct in the digital environment provide evidence that unauthorised file-sharing and other forms of uncompensated access to copyright works have progressively lost appeal due to the significant growth of lawful content platforms. Distribution models enabling users to access creative works in a smooth, cheap and secure way, without requiring them to download and store permanent copies, appear the main reason users are abandoning illegal sites and services. However, the proliferation of new file-sharing services based on increasingly elusive technologies such as cyberlockers and torrents (Renda, 2011) has constantly enabled a significant portion of Internet users to escape the control of the creative industries and to access copyright works paying no remuneration whatsoever. Online piracy has remained significant because of the evolution of technologies that, taking advantage of cloud-based services, have made large amounts of unauthorised copyright works accessible to the public. For instance, one of the most significant mechanisms enabling large-scale distribution of unauthorised works is that of cyberlockers, used by popular services such as Megaupload (available until 2012) and Rapidshare to store and access content through servers located in jurisdictions where online copyright enforcement is out of reach (so-called “copyright havens”). These services, together with other platforms offering illegal streaming services, work exactly as licensed platforms such as iTunes, Spotify and Netflix. Some of these services also provide devices facilitating copyright infringement, such as decoders or set-top boxes that can be plugged into TVs, with add-ons containing links to websites enabling access to free and unauthorised streams of copyright-protected movies, TV programs and series, music and games. Other services such as “torrent” sites (The Pirate Bay is the most famous example) use peer-to-peer technology to enable site-based downloads of copyright works in dispersed segments of data which are later reassembled after having been indexed and categorized. Interestingly, in recent cases the Court of Justice of the European Union found that activities helping users access unauthorised copyright works can be viewed as acts of communication of the works to the public, which directly infringe upon the related rights of film producers.7

3.1.2. Uncertainties about the economic impact of piracy on remuneration of creators

File-sharing and online piracy have not only raised questions about the effect on sales and economic harm suffered by content creators in sectors (music, films, TV programs, video games) targeted by peer-to-peer communications and
copyright-infringing services. The phenomenon has also triggered a debate on whether unauthorised dissemination of copyright works might have also positive effects for creators. Statistically, it is undisputed that, in the music sector, which was hit first and directly by piracy, global revenues from physical and digital music sales declined by 42% between 1999 and 2014 (from $25.2 to 14.6 billion) (IFPI, 2018). It was only in 2015 that music sales started growing again. In 2017, which was the third consecutive year of growth, global revenues amounted to $17.3 billion (IFPI, 2018). However, it still contested whether unauthorised copyright works have become substitutes for purchased contents, fared access to online services or cinema visits. For instance, recent studies have found no evidence of digital music sales displacement, reaching the conclusion that Internet users do not view illegal downloading as a substitute for legal access to digital music (Aguiar and Martens, 2013; Frosio, 2016).

From a broader economic perspective, economists have observed different interactions between piracy and sales. It has been observed that, by exposing consumers to music, film, books and games (and to artists, authors and genres), piracy has had a sampling effect that created new demand (Fijk, Poort and Rutten, 2010). This demand has also enhanced consumer willingness to pay for complementary products, such as concerts and merchandise, that have benefited the music industry as a whole. It is evident that these effects vary significantly depending on the type of works: for instance, the sampling effect is stronger with regard to works, such as sound recordings, that consumers tend to enjoy many times; it is much weaker for works, such as films and books, that are viewed and read once or twice. Also when it comes to complementary products, it is evident that musicians who gain popularity and exposure from unauthorised, free access to their recordings and videos have more to gain from piracy than non-performing artists, such as film directors and visual artists, who cannot earn money from their live performances. It is also relevant to consider that the positive effects of piracy in terms of online exposure gained by the artists can be greatly reduced if Internet users have access to copies of their works (for instance sound recordings or videos) whose quality is bad or compromised (for instance a badly compressed audio file or a bootleg video file) and/or whose formats come with no mention of their names. In both cases, unauthorised online dissemination is not only a violation of creators’ economic rights but, in those jurisdictions that protect moral rights, an infringement of creators’ rights to integrity and paternity.

3.2. Content-sharing platforms

In an October 2013 Guardian piece, the former leader of the Talking Heads, David Byrne, was very pessimistic about how the Internet would impact the commercial value of copyright (Byrne, 2013): “The Internet will suck all creative content out of the world.”

3.2.1. Platforms’ value and revenue generating strategy

Byrne’s voice has not been isolated in emphasizing a situation that is due not only to the notoriously weak bargaining power of the average authors and performers vis-à-vis content producers (e.g. record labels, film studios, etc) but also to the widely uncompensated dissemination of copyright works on content-sharing platforms. For instance, in a letter addressed to the European Commission’s President, Jean-Claude Juncker, signed in July 2016 by almost 1300 artists and songwriters from across Europe or who regularly perform in Europe, the artists claimed that the future of music was jeopardized by a substantial “value gap” caused by user-upload services, like Google’s YouTube, that were taking value away from the music community and from its artists and songwriters (IFPI & IMPALA, 2016).

The value and number of users of services such as YouTube ($70 billion in 2015), Pinterest ($12 billion) and Soundcloud ($700 million) easily evidence the central role and size of content-sharing services in online distribution of creative works (European Commission, 2016). As of October 2015, YouTube has 1.3 billion users (one third of all Internet users) who collectively upload 400 hours of video content every minute; Daily Motion has 300 million users watching 3.5 billion views every month; Vimeo has a monthly audience of approximately 170 million users and 35 million registered users; and Soundcloud’s user community has grown exponentially, going from 11 million users in 2011, to 150
Dr Giuseppe Mazziotti

Remuneration of Content Creators in the Digital Space:

million in 2015 and 250 million in 2016. More than ten years after the first publication of a YouTube video (2006), however, there is still uncertainty about the conditions under which content-sharing platforms are legally obliged to remove or filter unauthorized transmissions of copyright works and when they become liable for copyright infringement. First, this situation of uncertainty has not encouraged the development of licensing agreements between rights-holders and the enterprises owning the platforms. Second, as pointed out above, enforcement opportunities end up being reserved for the initiative of those rights-holders who are in a position to monitor user uploads on online platforms and to promptly notify online intermediaries.

What is the core business of large content-sharing platforms? YouTube, for example, helps demonstrate how platforms work, at least when it comes to revenue generating strategy. After having started as a free platform for user-generated content, and after its acquisition by Google in 2006, YouTube became a platform to share copyright works and also where copyright has been enforced to a significant extent (Renda et al 2015; see infra section 4). YouTube is not an online store or a commercial service where consumers pay a fee to access content. Rather, its business model (much like Facebook’s) looks like that of traditional broadcasters, where money comes from advertisers willing to pay for consumer attention. However, unlike free-to-air broadcasters, platforms such as YouTube or Facebook have neither editorial responsibility nor an institutional mission to inform, educate and entertain. The fact that users create or choose to upload all contents makes it simply impossible for platforms to guarantee diversity of accessible works.

3.2.2. Position of content creators on content-sharing platforms

From a legal and commercial point of view, the vast majority of user-generated content platforms (with the remarkable exception of YouTube in the last few years) and social networks do not easily ensure compliance with copyright. These platforms have given rise to a ‘lose-lose’ situation for content creators, in particular for individual authors and performers and small-size content producers:

- First, copyright holders have not been able to enforce their rights when a third party makes available their works without permission. If an author or a copyright holder has no resources to monitor user uploads and to send notices to take unauthorised contents down, their copyright will remain ineffective unless the platform deviser acts spontaneously (relying on content identification technology such as Google’s Content ID) and removes the unauthorised work. The above-mentioned “notice-and-takedown” procedures are mostly used, in the realm of content-sharing platforms, by major players in the music and film and by their respective anti-piracy bodies, in which these industries have invested significant amounts of money. As pointed out above, in the absence of active cooperation from the platform devisers, these procedures are not effective for individual content creators (such as photographers, writers, composers and film or video makers) and small content producers who do not have time and resources to dedicate to online enforcement. Thus, the current enforcement system discriminates against average content creators and favors wealthy ones.

- Second, the standard ‘Terms and Conditions’ that users of content-sharing services are normally required to accept at the time their accounts are created give the service provider (for instance Facebook or Instagram) a global, free and perpetual licence for the provider to use and exploit all user-authored contents across the platform, on a territorially unrestricted basis. This means that content-sharing services impose a condition of gratuity of use on the authors of available content. This means also that acceptance of the platform’s (non-negotiable) terms and conditions make content creators instantaneously lose their opportunities to be remunerated across the platforms, unless the service provider allows the account holder to monetize their successful content. Moreover, one must consider that the condition of gratuity in publishing creative works across platforms is not necessarily justified and made easier to accept by the remarkable exposure opportunities offered by the largest services. As argued above with regards to the sampling effect of music piracy, the only content creators who can view uncompensated uses or viewings of their works as a way to boost their live performance businesses are performing artists. This is
clearly not a publicity mechanism that works for non-performing artists, such as authors of films, videos and other audiovisual works, photos and journalistic content, for which unpaid dissemination via social media does not necessarily boost their chances to be remunerated.

3.3. On-demand content services

From the perspective of diversity of content made available to the public, the functioning of licensed content platforms exacerbates pre-existing inequality of success and distribution of copyright works as well as disparities in income among different authors, works and repertoires. Nonetheless, for works such as films and TV shows, a predominantly territorial model of online deliveries, for both download and streaming services, facilitates access to culturally and linguistically diverse content.

3.3.1 Inequality of success and income in digital markets

As observed in relevant literature (Taleb 2007; Renda et al 2015) authors and performers work in a scalable and very unequal environment where very few superstars have a disproportionately high share of the market, while the majority earns below average income (Towse, 2018). In larger and larger digital markets for creative works, scalability is induced by the “winner-takes-all” nature of success, combined with self-reinforcing trends. Moreover, even though the remuneration these services paid to a given licensor and to a group of right-holders can be identified, it is hard to assess how much creators gain concretely. This is because copyright licences normally contain non-disclosure clauses that allow service suppliers to keep such information secret.

This situation of opacity is even more complicated for works, such as musical compositions, for which online services negotiate fees and conclude agreements with collecting societies that manage the rights of thousands of composers and lyricists. The fact that, so far, these bodies have not (or not always) guaranteed an efficient and transparent use of the detailed information they need to identify their repertoire and monitor its effective use has not allowed a fine-grained (and, eventually, fair) allocation of revenues. Even though opacity and lack of a nuanced and technologically advanced management of all authors’ rights penalise mostly authors and/or owners of niche and small repertoires, music stars do not seem to earn so much from streaming services either. In the above-mentioned article Byrne mentioned the example of the 2013 summer hit “Get Lucky” by Daft Punk, where the two authors of the song (and members of the band) earned approximately 13,000 USD each as a result of the 104,760,000 Spotify streams this track reached by the end of August 2013. Byrne asked, “what happens to the bands who don’t have international summer hits?”.

3.3.2. Diversity of works made available

Due to the technically borderless dimension of the Internet, in online markets the role of physical distance between consumers and the place where digital content is made available to the public has sharply diminished. However, due to geographical restrictions implemented by platforms such as Apple’s iTunes, a 2015 study showed that less than a half of all songs and music albums were simultaneously available across the EU via Apple’s national music stores (Gomez and Martens, 2015). The study also found that – because of commercial strategies that draw on drivers of content demand such as language and home market bias – music availability was somewhere between 73 and 82% of what it could have been in a unified, unrestricted market. The situation was even worse for digital movies, whose simultaneous availability was estimated at 40% of the whole amount of content made available by iTunes in the whole EU. In spite of potentially global audiences, on-demand content suppliers such as Apple, Amazon, Netflix and Hulu still deliver films, TV shows and other audiovisual works produced by third parties as well as major sporting events on a strictly territorial basis. In the same way as free-to-air and pay-per-view TV broadcasters, these online platforms are bound by licensing agreements that establish areas of absolute territorial exclusivity, with a subsequent obligation
for the service providers to geo-block their signals and online transmissions (Mazziotti, 2019). Interestingly, given their potential to partially replace television broadcasting and compete for the same audiences, in certain jurisdictions these web-based services have started being subjected to some of the traditional duties of licensed broadcasters. In a culturally and linguistically diverse environment such as the European Union, for instance, video-on-demand (VoD) service providers are obliged under a recently amended media law directive to promote production of, and access to, European audiovisual works. This means that enterprises such as Amazon and Netflix will have to provide financial contributions to European content production (as determined by each EU member state), include a minimum share (30%) of European works in their catalogues and ensure prominence of those works in their offerings.

4. What has been done so far? Responses from governments

This section considers how governments have responded to challenges triggered by digitization of cultural content, unauthorised sharing of copyright works and the goal to preserve cultural diversity in an environment where economic characteristics of the digital market exacerbate inequality in content distribution.

4.1. International treaties and their broadly protective approach to copyright

At international level, multilateral instruments such as the 1994 TRIPS Agreement, the 1996 WIPO “Internet” treaties on copyright and the related rights of music performers and record producers as well as the more recent 2012 Beijing Treaty on audiovisual performances built up or consolidated a broadly protective system of intellectual property rights. In particular, the incorporation of the most important and comprehensive agreement protection of literary and artistic property — the 1886 Berne Convention (last revised in Paris in 1971) — into the TRIPS Agreement and, as a result, into the law of the World Trade Organization (WTO), strengthened reliance of the global economy on copyright. From an historical and legal perspective, the Berne Convention created a bridge between different copyright law traditions and aimed to establish international minimum standards, obliging its contracting parties to protect authors’ economic and moral rights on condition of reciprocity. The extension of the binding effects of the Berne Convention to all the members of the World Trade Organisation (WTO) not only introduced a relevant author-centric approach in an eminently corporate-related framework; it also made the obligations of the Berne Convention more easily and effectively enforceable against states through the arbitration-based dispute resolution system of WTO law.

Today’s international copyright system is the result of agreements that, from the Berne Convention onwards, created a multi-layer protection of individuals authors, performers and cultural industries (such as film and record producers and TV broadcasters). Most of these agreements are administered by the World Intellectual Property Organization (WIPO, which is a United Nations agency) and require the contracting parties to protect the work and investments of the most significant contributors to value chains of content creations through exclusive or remuneration rights. Since the mid-1990s, all the aforementioned individuals and entities have held broad exploitation rights whose scope covers, in addition to offline uses, all the economically relevant forms of copying and interactive distribution of content via the Internet and other digital means.

4.2. Copyright’s territoriality

In spite of the international origin and character of a vast array of today’s copyright norms, these rules do not automatically apply to copyright-related activities occurring on the Internet. These provisions need to be transposed, enforced and complemented at a national level by single states or regional lawmakers, such as Canada, the United States or the European Union, each of which has a legislative history, a copyright tradition and business and contractual practices that might vary from each other, even significantly. This means that creators’ rights are enforced on a country-by-country basis and the above-mentioned international norms are reflected in the laws of each single jurisdiction. The fact that the copyright system follows a principle of territoriality, even in a technically borderless digital environment, means
that national governments (and courts) are expected to fill the gaps left by the international conventions and to develop their own cultural, industrial and technology policies and laws, each of which can give the exercise and enforcement of copyright a different connotation. For instance, international conventions do not define what a copyright “work” is and when a work is “original”; convention rules do not indicate who should be considered to be the author (or coauthor) of a film and how the related rights should be allocated; and national law makers are also free to determine how the rights codified at international levels should be transferred from original right-holders to third parties (copyright contract law, see infra 4.3.1) and how such rights can be enforced in offline and online settings.

4.3. Copyright and freedom of contract

The copyright system is based, to a large extent, on a principle of free transferability of rights. Especially in the Anglo-American copyright tradition, freedom of contract is crucial in the creative industries. Such freedom is so broad that the US Copyright Act contemplates a category of works “made for hire.” Under US law, a work is made for hire when it is created by an employee within the scope of his or her employment or is specially ordered or commissioned for certain uses. In these cases, an employer or a client is considered the author even though the work is created by someone else. In legal systems following the French tradition of droit d’auteur (“author’s right”), instead, parties do not have such freedom. The notion of authorship under continental-European copyright systems entails also the acquisition, at the time of creation, of non-waivable moral rights to paternity and integrity that exist together with the author’s rights of commercial exploitations. In those systems, a work “made for hire” agreement would be contrary to the concept of authorship as a personality right and would be inevitably null or void.

4.3.1. Copyright contract law and revocation (or termination) rights

In jurisdictions which do not follow the Anglo-American “market knows best” model, the intersection of copyright and contract reveals a “paternalistic” approach aimed at protecting individual authors or performers in markets (for instance, book and music publishing) which are very risky and volatile. The policy goal of these measures is that of not allowing excessively lengthy or imbalanced transfers of rights in order to mitigate conflicts between individual creators and publishers. This is a clear attempt to remedy under national laws a situation in which copyright is concentrated, for the most part, in the hands of the cultural industries. Good examples of limitations of freedom of contract can be found in countries where the law seeks to help individual authors benefit from the economic value of their works at a time when their commercial success is very uncertain or impossible to estimate. For instance, in countries such as Germany, the Netherlands, France and Spain there is a principle of fair or adequate remuneration whose concrete determination is achieved also via collective bargaining (Germany) or through a government intervention (for instance, the minister of education, culture and science in the Netherlands), or on a sector-by-sector basis (Dusollier et al., 2014; Senftleben, 2018). This principle restricts parties from stipulating a transfer of the author’s rights in exchange for a lump sum and oblige parties to share market risks. In civil law jurisdictions such as Italy, France, Belgium, Germany and Spain parties are not free to transfer rights in future works of an author or in future modes of commercial exploitation. Moreover, in most of the above-mentioned countries specific types of transfers are regulated in depth. Publishing contracts, in particular, can be rigidly regulated in terms of formalities, duration and non-waivable rights of writers to have access to information concerning the revenues generated from the exploitation of the work, the quantity sold and the rights transferred for each exploitation of the work (see for instance Belgian and Italian law) (Dusollier et al., 2014). Publishers are also obliged by law to ensure distribution of the work to the public and, if they do not do so and let a book go out-of-print, they can lose their rights because of termination of the contract by law (see French law, for instance). All the aforementioned rights place authors and artists in a position to take advantage of the commercial success and diffusion of their works and can be particularly relevant in the digital environment.
A creator’s right that is more common across a variety of jurisdictions is the right to revocation (or termination) of contractual transfers. US law has conferred this right to authors following the 1976 reform of the US Copyright Act. As explained in the literature (Ginsburg, 2018) the US non-waivable author’s right to termination of the grant of their rights replaced a previous system (of “reversion”) where rights were automatically reassigned to the author after expiration of the first term of copyright protection (this right was based on a two-term copyright system). Given that the 1976 reform introduced a unitary copyright term — in order to make US law comply with the minimum term of protection required by the Berne Convention (life of the author plus fifty years) — the Copyright Act replaced the reversion right with a termination right. The main difference is that the new right is not automatic and requires the author, after thirty-five years from the grant, to properly notify the grantee and to record the notification in the copyright office within the statutory deadlines. Another relevant difference is that the termination right restricts freedom of contract because the right is enforceable “notwithstanding any agreement to the contrary,” whereas the previous right to reversion upon renewal could be overridden through contract by publishers (Ginsburg, 2018). Canadian law also creates a contractually non-waivable reversion right, even though it is provided in favour of the author’s heirs for them to regain and enjoy the author’s rights upon his or her death. To a more limited extent, termination rights are also applied in European countries, where they are granted in the context of regulated types of agreements. In France and Spain, for instance, the law applicable to publishing contracts grants authors a termination right should the publishers fail to fulfil their obligations. In the UK, instead, although the law no longer provides reversion rights (it used to do so until the creation of the 1956 Copyright Act), an automatic reversion of rights can be agreed contractually while transferring copyright.

4.3.2. Non-waivable remuneration rights for authors and performers

In the digital environment, EU law provides authors and performers with non-waivable rights to remuneration applicable to private copying of phonograms, audiovisual works and exploitation of music performances embodied in sound recordings. These measures guarantee a given income to individual creators with the intent to support their artistic career and/or to protect their financial interests. The impossibility of contractually relinquishing the remuneration rights listed below is important, from a policy perspective, since it aims to directly support copyright holders — the creative individuals — who are expected to guarantee diversity of cultural creations more than content producers and/or other commercial agents.

- Under EU law, private copying is a copyright exception whose legitimacy depends on fair compensation the right-holders receive for the economic harm they suffer from unauthorised copying. Technically, such fair compensation comes from “taxes” (“levies” in legislative jargon) the EU member states charge on the sale of copying devices such as printers and blank media or storage devices (including digital equipment like music and video game players, tablet computers, mobile phones, etc.). As an alternative to levies, this remuneration may come from state funds, like in Norway. This source of income — which is managed by authors’ and performers’ collecting societies — has been very important in Europe, especially for performers, since in most EU jurisdictions law provides that this revenue cannot be validly relinquished or transferred through contract to other right-holders like book and music publishers, film producers and TV broadcasters (Mazziotti, 2013).

- A 2011 EU directive extended the term of protection for sound recordings from 50 to 70 years from the time of publication or communication to the public (whichever is earlier). The measure was intended to provide additional revenues and economic support to both record producers and music performers. Considering that performers usually transfer or assign their rights to record producers, there would have been a risk of performers also contractually relinquishing this additional income. To solve this issue, the directive grants performers a non-waivable right consisting of a supplementary remuneration of 20% of the net annual revenues the record producer derives from exploitation of a recording in the extended term. For performers remunerated through lump sums, EU law obliges
remuneration of content creators in the digital space: dr giuseppe mazziotti

record producers to create a fund through which they distribute royalties on an annual basis via the administration of collecting societies. for performers remunerated through royalties, instead, the directive extends the right to receive such recurring payments, unencumbered by advance payments or contractually agreed deductions, during the extended period of protection. the directive also grants music performers a termination right when the record producer does not effectively market a sound recording during the period of extended protection.

4.3.3. specificity of the labour market in the arts sector

All these rights reveal a protective approach of the individual creator whose goal can be compared to that of labour laws. However, as economists have shown, the labour market in the arts sector differs from other labour markets because of an excess supply, due to too many artists being trained in colleges and academies and ending up not finding the type of job they hoped to do (towse, 2018). Studies have shown that only a portion of performers’ income is due to “arts” work, whereas the rest of their income comes from arts-related (for instance teaching) or unrelated occupations (baumol and bowen, 1966). This means that measures aimed at achieving fair remuneration of performers relate only to the artistic portion of their work. for instance, available data shows that the revenues generated by levies, in Europe, amount to 31% (2017) of the annual remuneration performers receive from their respective collective rights management organizations (which administer also the performer rights to equitable remuneration for broadcasting and communication to the public: aePO-ARTIS, 2018).

4.4. empowering creators through transparency of information

Policy makers have been trying to strengthen authors’ and performers’ bargaining power to help them increase their remuneration, especially from online exploitations of their works. a relevant attempt that is being made by the European union in adopting a new copyright directive for its “digital single market” is that of targeting the above-mentioned opacity of conditions under which creative works are accessed and licensed on social media, user-generated content platforms and on-demand services such as streaming platforms and online stores.15 the draft directive creates a right for authors and performers to receive, on a regular basis, timely, accurate, relevant and comprehensive information on modes of exploitation of their works, direct and indirect revenues generated, and remuneration due. This right to transparency will be actionable, via voluntary dispute resolution procedures, not only against content producers — which are contractual partners of authors and performers — but also against further licensees or assignees, including the owners of online platforms which buy online exploitation rights from collecting societies (in the music sector), and record and film producers and broadcasters (with regard to tv content). in the architecture of the upcoming copyright directive, transparency is functional to the enforcement of newly codified authors’ and performers’ rights, each of which will trigger a significant reform of national copyright and contract laws (see supra 4.3.1). the first is a right to contract adjustments when an author or performer’s remuneration is disproportionately low when compared to the subsequent relevant direct or indirect revenues deriving from exploitation. the second is a right to revocation of licences or transfers of copyright where there is an absence of exploitation of the work or there is a continuous lack of reporting of information on revenues and the remuneration due. as shown above (supra 4.3.1), both rights already exist in several EU member states. the upcoming directive aims at making them mandatory on a EU-wide basis.

4.5. collective management of authors’ rights in the music sector

Governments have sought to empower content creators by improving the governance and functioning of collecting societies, in order to ensure a more efficient, transparent and quick response to the challenge of newly emerged web-based uses of music repertoires.
4.5.1. Collecting societies as composers’ unions

Traditionally, collective rights management organizations are associations that work as unions, helping authors solve conflicts arising with music publishers (Mazziotti, 2011). One of the historically most significant achievements of collecting societies in the music sector has been to allow authors, through collective bargaining, to keep and co-own with publishers, on a fifty-fifty per cent basis, the rights these societies administer. This means that — unlike music performers — music authors (composers and lyricists), because of their membership in collective organizations that manage their rights on the grounds of a mandate, have never transferred their rights to music publishers in their entirety. This deal has clearly protected their right to fair remuneration and allowed composers to earn more, sharing commercial risks with their publishers.

4.5.2. Types of rights managed collectively

The rights these bodies traditionally manage are — according to a pre-digital subdivision of trade in two separate sectors — mechanical rights and performing rights. Mechanical (or reproduction) rights cover production and distribution of physical formats embodying musical compositions (for instance compact discs). Performing rights are much broader and target concerts and other public performances of a copyright work as well as transmissions via TV and radio broadcasting. In spite of the blurred distinction between copying and transmission of works over the Internet, collecting societies have maintained and relied upon this distinction in their online licensing activities. Mechanical and public performance rights have been transposed and applied to online uses to cover, to a different extent, both download and streaming services. Having said this, it is important to recall that authors’ societies in Anglo-American systems — unlike their continental-European sister societies — emerged and historically developed for the sole management of performing rights. In the UK, for instance, music publishers have historically been the sole proprietors of mechanical rights through their own trade organisations, after having acquired them from the authors. In continental Europe, instead, authors and music publishers usually co-own the same rights under the shield of their respective collecting societies (Mazziotti 2011). These bodies have operated on a strictly national basis and, in the vast majority of countries, they are de facto or legal monopolies which cooperate with each other, covering the entire global music repertoire, through mutual representation agreements.

4.5.3. Examples of remedies to inefficiencies and opacity

Unfortunately for authors and the entire creative sector, collecting societies were widely unprepared and slow to launch licensing schemes for online uses at a time when file-sharing and piracy dramatically affected the music industry and lawful content services struggled to emerge. In 2005, the European Commission set out best practices with a goal of enhancing efficiency and transparency of these bodies and encouraging them to provide better services to their members and to potential exploiters.16 These practices concerned crucial aspects such as equitable royalty collection and distribution without discrimination on the grounds of residence, nationality or category of the right-holders; increased collective rights managers’ accountability; fair right-holders’ representation in the organization’s internal decision-making; and effective dispute resolution procedures. After years of reluctance to intervene through mandatory provisions in a sector where national governments wanted to preserve their autonomy and cultural policies, these best practices were implemented in a 2014 directive that established a common legal framework for collecting societies in Europe.17 This directive allows authors (composers and lyricists) to entrust their rights to a society of their choice, irrespective of their country of residence, and to split the assignment of their rights between different societies (Article 5). At the same time, the directive (Title II) harmonised the criteria of governance and the main obligations of collecting societies, imposing high standards of transparency and fairness towards right-holders and commercial users of copyright works. With specific regard to online licensing of music rights, the directive is relevant because of the high standards of service and the technical requirements (for instance, use of time-sensitive and authoritative databases, processing usage reports and invoicing) it imposes on societies wishing to issue licences for cross-border online uses.
4.6. Online copyright enforcement

Facing exponential growth in uncompensated access to online copyright works, governments have attempted (often unsuccessfully) to reduce piracy to an acceptable level by undertaking initiatives at both international and national level.

4.6.1 International anti-piracy policy initiatives: failure of the Anti-Counterfeiting Trade Agreement (ACTA)

The most remarkable attempt to develop a coordinated response to the problem of piracy through a multilateral instrument was the negotiation and conclusion of the November 2010 Anti-Counterfeiting Trade Agreement (better known with its acronym 'ACTA').\(^{18}\) The treaty was an attempt to establish international standards and common rules to tackle large-scale infringements of all intellectual property rights (copyrights, trademarks, patents, designs and geographical indications). According to its supporters, ACTA would place intellectual property right-holders in a position to benefit from improved access to justice, customs, and police to enforce their rights against counterfeiters or infringers in all the countries where the agreement was entered into force.

Unsurprisingly, the treaty had a provision (Article 27.2) targeted at large-scale copyright infringements. ACTA sought to make civil and criminal enforcement measures contemplated in the agreement available also in the digital environment. This provision encompassed precautionary measures (such as injunctions) aimed at preventing and discouraging infringement. In particular, the provision targeted “the unlawful use of means of widespread distribution for infringing purposes.” In its final version, Article 27.2 provided that copyright enforcement measures should be implemented in a way that does not conflict with legitimate activities, including electronic commerce, and preserved fundamental principles such as freedom of expression, fair process and Internet user privacy. In spite of amendments that sought to explicitly strike a balance between copyright enforcement and competing fundamental rights, the agreement could not be formalised after its unexpected rejection by one of the most relevant Contracting Parties, the EU. Even though the EU and twenty-two governments of its member states signed ACTA in January 2012 (one month after the other parties) the European Parliament rejected the treaty ratification in its consent procedure, in July 2012, after an unprecedented and politically harsh debate and widespread protests across Europe. In 2010 the European Parliament had already openly contested, through a formal resolution, ACTA’s lack of transparency in the negotiations, asking the European Commission (which acted as the EU negotiating body) for an assessment of the potential impact of the new treaty on freedom of expression and the user right to privacy.

4.6.2. Examples of national anti-piracy policies: HADOPI and SOPA

One of the main reasons for the conflict between the European Parliament and the European Commission during the ACTA negotiations was the inclusion in the original text of the treaty of a provision that encouraged the adoption of a so-called “graduated response” (or “three-strikes”) law. This enforcement model was followed in the ‘Creation and Internet’ Act adopted in France in 2009.\(^{19}\) This act established the HADOPI (Haute Autorité pour la Diffusion des Oeuvres et la Protection des droits sur Internet), an administrative agency with the institutional mission of sanctioning Internet users accused of illegal file-sharing. Under the original version of the “three-strikes” law Internet service providers were requested to monitor infringing conduct by their subscribers and, after three warnings, to place them in a blacklist and to block their account for up to one year (with the “three-strikers” continuing to pay while being disconnected: Renda et al., 2015). As argued in the literature (Renda et al., 2015), this piece of legislation violated the principle of net neutrality by requesting the Internet service providers to monitor online traffic and to detect copyright infringement. What made the French model law very controversial and unacceptable for the majority of the European Parliament members was also the fact that an administrative body such as HADOPI — and not a judicial authority — could sanction users and order disconnection of their accounts.
from the Internet. These concerns were fully reflected in a 2009 judgment of the French Constitutional Council which censored the new law because of its inconsistency with the presumption of innocence and the right to a fair trial. This judgment forced the French parliament to amend the act and to confer the power of cutting off Internet access of repeat copyright infringers to courts.

Another interesting example of national anti-piracy law initiative that failed in January 2012 because of fierce opposition coming from Internet companies and civic society organizations was the Stop Online Piracy Act (SOPA). As its title suggested, this act aimed to expand the ability of the US law enforcement authorities to fight online piracy. The US legislative initiative contained specific provisions enabling courts to issue site-blocking orders targeted at Internet service providers and other judicial remedies that could have restricted advertising networks and payment services from conducting business with infringing websites and prohibited search engines from linking to such websites. Interestingly, EU law already contemplates site-blocking measures, to such an extent that member states such as Ireland and the UK have made them particularly effective against entire websites and services engaging in large-scale copyright infringement. Given the evolution of piracy and its significant shift to the Cloud and to websites implementing sophisticated peer-to-peer technologies, major film producers and audiovisual content coalitions such as FairPlay Canada have lobbied governments and communications authorities to access these measures and impair or reduce access to structurally infringing sites (for instance, The Pirate Bay).

4.6.3. Online platforms’ liability exemptions

From the late 1990s onwards, governments enacted regulations to shield providers of certain web-based services (including content hosting) from liability for activities carried out by their users (see supra 2.2.2). The widely shared policy goal pursued by these measures was to foster the development of a solid digital communication infrastructure, encouraging innovation and maintaining a principle of network neutrality. However, there have been significant differences in the ways these liability exemptions have been applied to content-sharing platforms in the European Union and the US.

In Europe, the EU Court of Justice (L’Oréal v. eBay, 2011) stressed and clarified that a liability exemption is applicable to online platforms in so far as a platform confines itself to providing a hosting service neutrally, by a merely technical and automatic processing of the (potentially infringing) contents uploaded by its customers. This means that the exemption should not apply when an online intermediary plays an active role that entails knowledge of (or control over) such content. For instance, the CJEU found that this was eBay’s role in supplying assistance and optimising presentations of the customers’ sale offers or promotion of these offers. In the domain of content-sharing platforms, this means that the service provider cannot escape copyright liability if it optimises the presentation of the uploaded works or promotes them. This is the conclusion that might soon be codified in an EU directive on copyright in the “Digital Single Market.” The draft directive contains a provision (Article 13) that explicitly considers user uploads on content-sharing platforms as acts of making works available to the public. The same provision obliges providers of such services to seek and obtain licenses for copyright works that platform users, acting for noncommercial purposes, share online.

In the US, instead, the 1998 Digital Millennium Copyright Act and its safe harbour provisions have a broader application that encompasses and covers almost any Internet entities (Ginsburg and Budiardjo, 2018). US courts have recently held that video-sharing platforms such as YouTube and Vimeo can seek safe harbour when they prove absence of knowledge or awareness of facts or circumstances from which infringing activity is apparent (so-called “red flag” knowledge). This approach is motivated by the intent to protect Internet service providers from the expense of monitoring user uploads, which was a specific concern of the US Congress when designing the safe harbour provisions.
5. Responses from online platforms and civil society

Proliferation of user-generated content (or “content-sharing”) platforms and social media has empowered Internet users’ creativity and has allowed creators to make their works available to the public. However, as emphasized above, for several reasons, these platforms have facilitated copyright infringement by letting users copy, transmit and modify unauthorised works. This section briefly explains how the platform owners and civic society organizations have reacted to the challenge of copyright enforcement.

5.1. Online platforms

The broad implementation of liability exemptions and notice-and-takedown mechanisms (especially in the US) has been beneficial to online platforms because it has allowed them to not enforce — de facto — clauses under their terms and conditions that would restrict users from publishing unauthorised copyright materials. Since their emergence in the online world, platforms have given access to user-generated texts, photos, music, videos and other content without paying remuneration to copyright holders whose works were uploaded without their permission (or without letting them monetize those unlicensed exploitations, at least initially). However, not all platforms have behaved in the same way. Considering the narrower scope of liability exemptions in the EU and the constant pressure content industries exert judicially, especially in the US, most platforms have started enforcing copyright. This is happening through implementation of technologies which enable platforms to detect and remove copyright-infringing materials and give content creators opportunities to monetize use of their works.

The first platform which developed a content identification and rights management technology was YouTube. Having transformed its platform into a formidable advertising machine, Google sought to shield YouTube from copyright liability claims by investing dozens of millions of dollars to launch “Content ID” in 2007. This technology enables right-holders to effectively monitor and manage their works on YouTube by automatically notifying the platform of user-uploaded videos containing their creative works. The functioning of this system is based on cooperation with right-holders, who deliver to YouTube reference files of works they own, metadata describing the content and the option to choose in advance what they want YouTube to do when Content ID finds an appropriate match. As Google explains in a November 2018 document, “the library of reference material in the system includes more than 80 million files of audio and visual content. YouTube then compares videos uploaded to the site against those reference files and automatically identifies the work and applies the rightsholder’s preferred action for that content” (Google, 2018).

At a much later stage, Facebook started following the same path by implementing a “Rights Manager” technology that helps all platform users managing a Facebook page prevent their unauthorised videos from being spread across the social network. The tool allows content creators to easily upload and maintain a reference library of video content to monitor, protect and specify permitted uses of each video. This technology gives account holders the possibility of earning and claiming a part (or the whole) of the advertising revenues generated by their videos and embodies a function (“Audience Network”) helping users develop the most profitable publishing strategies. The main difference between YouTube’s Content ID is that Facebook’s technology is not its own product. The solution is licensed by Audible Magic, a company maintaining content identification databases for multiple types of content, including music, live and past TV programs and movies. Audible Magic’s solutions are also implemented by other content-sharing services such as Daily Motion, Vimeo, Soundcloud and Tumblr.

The recent takeoff and advancement of content identification technologies and databases for music, film and TV content is good news for content creators since these tools increase their ability to monitor and boost their remuneration opportunities. This is also a significant improvement for platforms since they can more easily handle obligations descending from liability regimes and help creators monetize their works under conditions negotiated ex ante. These solutions are also likely to foster cultural diversity by enabling creators of all kinds — from the least to the best...
Remuneration of Content Creators in the Digital Space:  

Dr Giuseppe Mazziotti

known, from the smallest producers to the majors — to publish their works and repertoires and to measure their diffusion and appreciation across platforms. However, it has yet to be seen how effective these technologies are, especially in relation to types of works — for instance, photographs on photo-sharing sites such as Instagram or Pinterest — whose recognition and classification is, at least for now, very difficult or impossible.

5.2 Civil society

As noted above, civil society organizations have been active in Europe and the US in opposing and lobbying against legislative reforms in the field of copyright that would have tightened online enforcement and established usage restrictions with the cooperation of Internet service providers and online content platforms. The political rejection of ACTA in 2012 showed how influential and decisive unprecedented waves of protest and lobbying by individual users, civil society organisations and associations representing the computer and communications industries can be. Coincidentally, the protest against ACTA followed, by a few months, a political battle that took place when the US sought to pass the Stop Online Piracy Act in January 2012. At the time the SOPA was discussed in the US Congress, a mobilisation of Internet users, online intermediaries and providers of content-sharing platforms culminated in a “strike” (i.e., a switching off of approximately 7,000 websites, including top websites like Wikipedia and Reddit) that effectively raised awareness about the risks of the new law (for instance, tension with the basic functioning of the Internet and risks of online censorship).

From a completely different angle, civil society organizations have contributed to a general improvement of the copyright ecosystem by developing standard mechanisms of rights management which aim at facilitating dissemination of creative works in the digital space. One such example was in the open source movement (and in particular the Free Software Foundation), which since the early 1980s promoted a new way to create, improve and disseminate free software as an alternative to the marketing of commercial programs. The most prominent and influential example of such a contribution in the Internet age is Creative Commons, a non-profit founded in California in 2002. Creative Commons has conceived, made available and constantly upgraded a system of standard licenses addressed to all content creators who intend to make their works free to access and share on the Internet under a set of easy-to-understand, flexible and predetermined conditions that address all kinds of digital works. From a policy perspective, Creative Commons pursues the goal of opposing the far-reaching scope of copyright from the inside of the system, promoting dissemination of culture, entertainment and knowledge on public interest grounds. From a legal perspective, the organization encourages content creators to replace the traditional “all rights reserved” approach to copyright with a “some rights reserved” rationale. This model is particularly useful for nonprofessional (amateur-like) creators and other authors who, for various reasons, do not seek remuneration from dissemination and use of their works. For instance, certain authors who are already remunerated for their work via an employment contract (e.g. university professors or researchers) or performers who receive an artist grant (such as fellowships given to artists in residence in concert seasons or festivals) might find it more useful to freely publish their scientific articles, research papers or classical music recordings than monetising their work. The standard licences proposed by Creative Commons aim at helping authors and artists relinquish some of their rights in order to make their works easier to access, share and — in certain cases — reuse by follow-on creators. Creative Commons (‘CC’) gives content creators the possibility of choosing among several clauses, each of which is expressed graphically through a simple icon. Each of the clause combinations allows an individual creator to concretely determine how “open” their creative work is to users. As shown in the image below, the minimum common denominator of these licences is an ‘Attribution’ (‘BY’) clause, which protects the creator’s right to be credited when their work is copied, shared and re-used in the online environment. Other clauses specify whether the work is made available just for ‘noncommercial’ purposes (‘NC’) and/or for ‘non-derivative’ uses (‘ND’) and whether a modified version of the work (in those cases where the ‘non-derivative’ clause is not applied) should be shared under the same conditions chosen by the original creator (share-alike: ‘SA’).
Remuneration of Content Creators in the Digital Space:

Dr Giuseppe Mazziotti

From the perspective of professional content creators and cultural industries seeking remuneration from online exploitations of their works, initiatives such as Creative Commons' are important because their standardised licences help individual and commercial users distinguish free-to-use materials (comparable, in this regard, to works fallen in the public domain) from copyright works whose unauthorised sharing constitutes copyright infringement. These licenses pursue this goal not only from a legal and contractual perspective, but also encourage the incorporation of a machine-readable version of each licence into the files embodying creative works (through metadata and forms of content tagging). This means that the Creative Commons' attempt to “de-activate” the default protection afforded by copyright is a message targeted at both humans and computers that can significantly improve copyright enforcement and, in particular, the ability of platforms to filter or monetise unauthorised materials without impairing dissemination of unprotected materials and freedom of speech.

6. Going forward: major obstacles

There are major obstacles to the achievement of conditions that might spur fairer remuneration of content creators in the digital space and a broader dissemination of diverse cultural expressions. This section draws on the scenarios that have been described so far in order to identify the main issues policy makers and participants in a multi-stakeholders dialogue should consider.

6.1 Secrecy, lack of transparency and scant data on content value chains

Estimating the commercial value of digital works is difficult or impossible. Dematerialization and disintermediation of content and the subsequent advent and exponential growth of content-sharing platforms has dramatically weakened Internet users' appreciation of professionally created works. Business models that have emerged in the last decade via social networks, user-generated content platforms and licensed on-demand services have cut off or reduced the power of previous retailers and commercial intermediaries and changed the value chains of content distribution. A quick transition from markets where consumers purchased physical copies of creative works to services where subscribers pay to access collections of works (or to freely access large amounts of user-uploaded content) has placed online platforms in a unique position to exploit digital content. Considering their market power, large platforms can determine and impose the conditions and price of the works they make available in environments where a significant portion (or all, in some cases) of their profits come from advertising and exploitation of their users' personal data. As noted above, secrecy and confidentiality cover all the information on revenues generated by copyright works and the levels of remuneration paid to creators or to their collecting societies. It is impossible to estimate fairness of remuneration and diversity of content online without data showing what the preferences of Internet users are. As we have seen, in a recent legislative proposal the EU intends to ensure transparency along the value chains of digital content by obliging assignees and licensees of copyright works — including online platforms — to disclose allocation of earnings and to inform authors and performers about the revenues generated by their works, on a sector by sector basis. However, it has yet to be seen how realistic disclosure and processing of such a vast array of data is, especially in jurisdictions where authors and performers are represented by inefficient or technologically poorly equipped collecting societies and where national lawmakers might not be determined to place such a heavy administrative burden on Internet companies.
6.2. Absence of rights management information standards

Obliging online exploiters to disclose relevant data on revenues generated by copyright works, on a sector by sector basis, would not be enough to ensure fair remuneration of the generality of content creators. Effective monitoring and measurement of access to digital works presupposes the implementation of content identification technologies and repertoire databases containing all necessary information about the relevant rights as well as who owns and/or controls them. Unfortunately, at the moment there are no fully interoperable standards giving content licensors and licensees access to rights management information. Availability of such data would greatly facilitate the operative elements of licensing agreements and would promote the creation of a level playing field for all contributors to the content value chains in different sectors. An attempt of this kind was done in the music sector with the so-called “Global Repertoire Database” (GRD). This ambitious project aimed at creating a comprehensive database of the global ownership and control of musical works, openly available to composers, publishers, collecting societies, and commercial users of the global repertoires. The GRD would have enabled cost savings — by eliminating duplication in activities of data management and processing — and would have allowed a more efficient management of online works by lowering administrative barriers for companies wishing to distribute music online. Such an open, reliable and fully interoperable database would also have ensured a quicker and more efficient compensation to content creators. Unfortunately, despite the support and involvement of all the big music publishers and some of the digital players (including Google) who would have needed access to the data, the project failed in 2014 because of lack of financial support from collecting societies that would have ended up benefiting from the initiative without having contributed to it. PRS For Music (UK) and Swedish collecting society STIM, which formed a joint venture to work as technology provider, were the only societies involved in this project. This unsuccessful attempt shows that a proprietary approach to the development of a standard database might not be the right solution. An alternative might be obliging the music publishers and collecting societies to put their databases into the public domain in order to enable third parties to develop a database and then let the market decide what solution is the best.

6.3. Bargaining power, size and origin of the largest online platforms

Today’s largest online platforms are all owned by dominant and very resourceful tech companies headquartered in the United States. These four companies are often referred to as “GAFA” (Google, Amazon, Facebook and Apple) or “over-the-top” digital content suppliers. Each has a different business model and their online platforms are under increasing scrutiny all over the world. A major concern is that Amazon, Facebook and Google have and exert too much economic power to the detriment of consumers, suppliers or competitors. An influential legal scholar has recently argued that Google, Facebook and Amazon are a threat to democracy as they become bigger and bigger and — for this reason — they could be broken up under antitrust law (Wu, 2018). However, antitrust does not seem an effective remedy against the excess of corporate power of the tech giants in the US, even if they are regarded as monopolists. In a leading antitrust case, the US Supreme Court held that monopoly is an important element of a free-market system and is desirable because it induces risk taking that produces innovation and economic growth. Moreover, when it comes to merger control, courts have consistently applied a “consumer welfare standard” under which the US government is entitled to block a merger — such as Facebook’s acquisition of Instagram and WhatsApp — only if it can prove that the merger results in increasing prices for consumers. As pointed out in the literature (Wu, 2018), applying this standard in markets where large companies offer web-based services for free makes antitrust scrutiny impossible. A positive phenomenon that helps preserve competition is that the largest online platforms, despite of their different business models, compete with each other in a disruptive way, starting to offer services and products that are at the core of their competitors’ business. For instance, Google started operating an ultimately unsuccessful social network, Google+ in response to Facebook; Apple and...
Facebook are both investing heavily in technologies which improve online search, etc. This cross-market competition has the potential to reduce the power and influence these companies have on the market. In the European Union, instead, the fact that antitrust law can be used to sanction abuses of dominant position enables the European Commission, acting as the EU antitrust authority, to target anticompetitive practices with the aim to protect competitors and not only consumers. This approach is evidenced by the Euro 2.4 billion fine the Commission issued in 2015 against Google for having suppressed search rivals by denying equal access to its platform in the context of shopping offerings.

A different approach to antitrust and to potential restriction of anticompetitive conduct in the US and Europe shows why it would be difficult to develop a shared understanding of the need to regulate online platforms at international level. The European Commission has clearly shown its intent to establish a legal framework where the largest content-sharing platforms will have enhanced responsibilities and decisive roles in preventing, removing and keeping offline a broad variety of illegal content, including copyright-infringing materials. In the US, instead, a broad implementation of the DMCA safe harbour provisions and persisting reliance on platform neutrality shows a radically different policy. Needless to say, this also has implications for how online platforms remunerate content creators and support cultural diversity. The US is not only the GAFA’s country of origin but is also home to the most successful creative industries in the world, including the biggest (mono-language) movie industry. This explains why US policy makers have traditionally shown no interest in fostering cultural or linguistic diversity and embracing a regulatory model such as the EU’s or Canada’s in the field of media law. From a US perspective, the fact that some of the online platforms might encourage large-scale use of unauthorised works, with a subsequent decrease of value for copyright, is more than compensated for by the continuous growth of the technology sector and development of industry-led solutions which allow the creative sector to control and/or monetize its productions.

6.4. Risks and costs entailed by online copyright enforcement

Fair remuneration for content creators and stronger support for dissemination of diverse cultural expressions would require a more efficient system of copyright management and enforcement, made easily accessible to individual right-holders and small content producers. However, as the previous sections have argued, enforcement measures in the online environment have to be handled carefully in order to avoid impairing freedom of expression and communication, net neutrality and freedom to do business online. Site-blocking measures disabling access to copyright-infringing sites and content filters aimed at removing unauthorised works from content-sharing platforms entail the risk of “overblocking.” Such measures can easily end up suppressing speech and free communications by also targeting works and other materials that are in the public domain or made available under licenses such as Creative Commons’. This is why online platforms should continue to monitor the implementation of content identification technologies (such as Google’s Content ID or Audible Magic’s solutions) and provide redress mechanisms to users whose works and materials have been erroneously removed.

Making enforcement measures comply with protection of fundamental rights and civil liberties will be the main challenge for policy makers in the close future. This is particularly important not only when enforcement is “delegated” to platforms but also when the power to take site-blocking measures is exercised by administrative authorities. For instance, this is the case in Italy since 2014, when the Italian Communications Authority (AGCOM) started handling site-blocking requests to speed up and simplify enforcement procedures and to supplement slow and ineffective civil proceedings. Obviously — given the principle of copyright territoriality — courts still grant different types of enforcement measures and evaluate conflicts between copyright and human rights in different ways and according to distinct exceptions and defences. Even in legal systems which are highly integrated, such as those of the EU member states, copyright holders have no pan-European copyright enforcement measures at their disposal, in spite of the
attempt to develop a “Digital Single Market.” This means that, even in the EU, content creators still have to enforce their rights on a country-by-country basis, facing costs and difficulties that only major content producers can handle.

Last but not least, the strong reactions and lobbying of civil society organizations and the largest Internet companies against policy initiatives aimed at improving copyright enforcement show how difficult it will be, from a political point of view, to gather sufficient consensus on new and future-proof reforms, at both international and national level.

7. “Remuneration” and content “creators”: towards a shared understanding and a common language at international level

What is the role of remuneration in ensuring sustainability of content creation in the long term and who is a “creator” in the online environment?

7.1. Notion of “creation” and different kinds of contributors to content value chains

This paper has adopted a broad notion of “creation,” which includes the work of individuals or groups of individuals, performers, small and medium-size enterprises and major producers such as big record companies, Hollywood studios and TV broadcasters. Under this definition, even Netflix can be viewed as a content creator after having started producing and distributing across its global platform an increasingly broad variety of originally produced movies and TV shows. Even though remuneration is not necessarily the main goal of content creation, especially for individual authors and performers, all of these different categories of creators expect to be compensated when the result of their work is disseminated and exploited commercially. The creative sector, as a whole and as a network of distinct industries, would not be able to produce new works on an ongoing basis without adequate economic incentives and rewards. Even though the value chains of content creation vary significantly from sector to sector, each production system, from the simplest to the most sophisticated, is based on the idea that the final output of content creation will be remunerated in one way or another.

7.2. Function (and limits) of copyright law

Copyright is the area of law and policy where the need to ensure remuneration of creative labour has traditionally been addressed, with the progressive establishment — partially through international agreements — of distinct rights in favour of authors, performers, content producers and broadcasters. However, as illustrated in the previous sections, copyright protection is not, as such, sufficient to guarantee an adequate level of remuneration across value chains of content production. This statement holds true even more so at a time when allocation of copyright revenues depends on arbitrary and secret decisions taken by a handful of technology companies that impose prices and conditions of access to their platforms, for both content creators and their users/subscribers. Copyright systems are generally neutral when it comes to levels of remuneration of right-holders and do not guarantee a given income, especially to right-holders with a weak bargaining power. This is the case for average authors and performers or owners of works and repertoires with limited international appeal — in the music, film and TV sectors — who can easily go unnoticed in scalable markets where the “winner-takes-all” nature of success means that a very few works and authors have a disproportionately high share of the market. Moreover, freedom of contract for large content producers holding market power, such as the film and music majors or large book publishers, tend to better protect corporate interests than individual creators’ expectation to gain fair remuneration. This means that cultural industries gather as many interests as possible — often for the whole life of the copyright — whereas authors and performers transfer their rights in exchange for a lump sum or royalties. Finally, as the sections on online piracy and enforcement have noted, the rights granted under copyright law — in spite of their broad scope at international and national level — have become difficult to enforce and monetize in web-based settings where fully decentralised forms of content distribution or content-sharing platforms make it difficult or impossible for copyright owners to negotiate and earn a fee from commercial and non-commercial users of their works.
7.3. Suggested elements of a multilateral dialogue

Given the aforementioned conditions, what are the responsibilities of governments, stakeholders and civil society? How can a shared understanding and a common language on remuneration and long-term sustainability of cultural creation be developed, from a multilateral and multi-stakeholder perspective?

- First, a fruitful discussion could start from considering whether a reconciliation of the corporate and cultural aspects of copyright is a realistic scenario. International copyright treaties, and in particular the incorporation of the author-centric Berne Convention into the domain of international trade and of WTO law (i.e., the TRIPS Agreement), show the existence of a shared understanding and a common language on trade-related aspects of copyright and the need to protect economic rights and interests of copyright owners, who are normally business entities and not original creators.

- Instead, there is no common language on how to protect the single elements of each content value chain and, in particular, intellectual labour and its potentially very broad diversity, which is ultimately conferred to creative output more by individuals than by enterprises. At present there is no cultural exception under the TRIPS agreement and WTO law that preserves sovereign states' autonomy and freedom, in the long term, to derogate from liberalization and free market rules in order to provide financial support and to reserve a preferential treatment to local productions, especially in the audiovisual sector. This is a sector where the EU and Canada, at least for now, fund — directly or indirectly — their domestic film and TV content industries by imposing quotas and other content requirements to licensed broadcasters in their own territories. These measures (that the EU has extended to online film and TV services such as Netflix) ultimately protect local works and repertoires from the US audiovisual industries’ competition, in a genuinely protectionist fashion.

- It is still unclear how a predominantly trade-related approach to copyright enshrined in the most important and effective treaty on intellectual property — the TRIPS Agreement — can coexist with freedoms and rights that the 2005 UNESCO Convention on diversity of cultural expressions grants to its parties in pursuing their cultural policies (Macmillan, 2014). Unsurprisingly, the contracting parties of this Convention do not include the US. The Convention, which does not mention or take copyright into consideration, expressly provides contracting parties with the right to adopt measures aimed at protecting and promoting diversity of cultural expressions within their own territory. States are therefore free to give their own cultural industries opportunities, means and financial aid for creation, dissemination, distribution and enjoyment of domestic works and repertoires, also on the grounds of the language used for such activities.

- What will happen to content creators when much greater amounts of their copyright works will circulate digitally and revenues that authors, artists and producers still gain from physical markets and uses will shrink? In order to strengthen the bargaining power of individual creators, governments might consider obliging rights transferees and their licensees to disclose data that would give individuals the possibility of verifying levels of effective content remuneration in various creative sectors. A similar duty of data disclosure could be placed on authors’ collecting societies and other rights licensors or agents holding rights ownership information that third parties need in order to build up standard repertoire databases on a sector-by-sector basis. As this paper advocates, greater transparency is and will be a key factor in balancing contractual relations between authors, producers at platforms at a time when globalization and web-based content markets are exacerbating inequalities among works, repertoires and authors.

- To develop a shared understanding and a common language on remuneration of creators, governments and stakeholders might consider going back to the original purpose of the Berne Convention on literary and artistic property: protecting individual authors and their relationship with the fruit of their intellectual work. A multilateral discussion might start from considering that the Berne Convention has a genuinely author-centric approach and protects also non-economic (“moral”) rights of authors to paternity and integrity of their works. As we have seen in the Creative Commons’ examples, such non-economic rights matter also for non-professional or non-profit-seeking creators because of their function to protect personal identity and reputation in increasingly
complex digital settings. At a time when remuneration for average authors is endangered, it would be promising for a multilateral body or working group to seek a stronger and more intense protection of individual creators’ rights that might reflect also the rationale and values of the UNESCO Convention. To this end, this multilateral discussion might focus on successful models of copyright contract law and harmonisation of criteria and conditions of exercise of authors’ rights to contractual adjustments and/or to revocation or termination of their rights transfers.

- In reconsidering the importance and function of the Berne Convention and its centrality in the architecture of the TRIPS Agreement, it would be useful for the overall multilateral discussion to clarify what room for manoeuvre the Convention leaves to contracting parties in testing or implementing reforms, such the introduction of certain formalities. A manifestation of interest coming from authors in having their works effectively protected in the online environment might help to improve searches and enhance clarity on rights ownership. Moreover, such input from authors could make it easier to protect professionally created works by keeping them distinct from the endless amounts of non-professional and user-generated works made available on content-sharing platforms every day. However, the Berne Convention restricts its parties from making the coming-into-being and the enforcement of the author’s right conditional on formalities such as registration of the work (Ginsburg, 2010). What the Convention does not restrict, however, and seems a promising policy option is the introduction of a requirement under national laws which could make copyright transfers valid and enforceable as long as these contracts were recorded in a public register or database. As observed in the literature (Van Gompel 2014), if governments followed this route, they would greatly encourage and simplify creation and free accessibility of rights management information. Interestingly, for the purpose of this discussion, accessibility of public records of rights transfers would show who the authors are, who effectively exercise their copyright in a professional way and what are the economic conditions under which a right has been assigned and for how long.

- Finally, a dialogue with the largest online platforms — in particular Google and Facebook, which control the content-sharing services with the highest number of users — would help a multilateral group of policy makers identify best practices on content identification and takedowns (or “notice-and-notice” regimes), redress mechanisms to the benefit of Internet users and licensing and monetization criteria. Given that these huge content gatekeepers operate their services on a territorially unrestricted basis, this soft law, as well as industry-led solutions developed under the supervision of national governments, could easily be tested and spread out, potentially, beyond national borders and reach a global scale. This dialogue could also help develop a model law for creation of new safe harbours targeted specifically at content-sharing platforms.
Remuneration of Content Creators
in the Digital Space:

Dr Giuseppe Mazziotti

References

1 EU Fulbright Scholar 2018/2019; Emile Noël Global Fellow 2018/2019, Jean Monnet Center for International and Regional Economic Law and Justice, New York University; Assistant Professor in intellectual property law (on leave), School of Law, Trinity College Dublin.

2 In cases involving the liability of operators of peer-to-peer platforms, the Ninth Circuit Court of Appeals and the US Supreme Court found Napster and Grokster, two popular providers of file-sharing software, indirectly liable of copyright infringement, even though these companies did not store infringing materials on their servers. Napster was found to have given its users the means to infringe copyright while having specific knowledge of such infringements: see A&M Records v. Napster, 239 F.3d 1004 (9th Cir. 2004), hereinafter Napster. Grokster, instead, was found to have induced or encouraged direct infringement by advertising infringing uses of its technology or giving instructions on how to infringe, even though it could not be aware, from a technical point of view, of the infringing activities. See MGM Studios v. Grokster, 545 US 913 (2005), hereinafter Grokster.

3 US Digital Millennium Copyright Act (DMCA) signed into law by President Clinton on 28 October 1998, which amended the U.S. Copyright Act (see US Code, Title 17).


5 Copyright Modernization Act (S.C. 2012, c. 20).


7 See C-527/15 Stichting Brein v Jack Frederik Wullems (2017), where the defendant was distributing a device (‘Filmspeler’), to be plugged into TVs, with add-ons containing links to websites that enabled access to free and unauthorised streams of copyright-protected movies; and C-610/15 Stichting Brein v Ziggo BV and XS4All Internet BV (2017), where the infringing activities consisted of creating, maintaining and supplying a system based on peer-to-peer software, operation of a website dedicated to the downloading of files (The Pirate Bay) and acts of indexing and categorizing dispersed segments of data that, after having been re-assembled, gave access to copyright works. In both cases the CJEU found that the defendants had infringed the right of communication to the public.


11 Beijing Treaty on Audiovisual Performances (not yet in force), adopted on 24 June 2012, available at http://wipo.int/treaties (Canada is not a signatory of this treaty).


13 See the TRIPS Agreement, Article 9 (Relation to the Berne Convention).

14 Section 101 of the US Copyright Act (Title 17 of the US Code).


18 The parties of the agreement were Australia, Canada, the European Union, Japan, Mexico, Morocco, New Zealand, Singapore, South Korea and the United States. The treaty was negotiated and concluded outside the institutional and legal framework established by the World Trade Organisation (WTO). Considering that all parties are WTO members and, as a result, are bound by the TRIPS Agreement (whose Part III includes provisions on enforcement of intellectual property rights), ACTA would have been a ‘TRIPS-plus’ instrument (an agreement providing more stringent rules and obligations on enforcement among the contracting parties).


20 See H.R. 3261, the “Stop Online Piracy Act”.

21 See C-324/09, L’Oreal and Others v eBay International AG and Others (2011), par. 116-124.

22 Viacom v. YouTube, 676 F.3d 19, 39 (2d Cir. 2012).

23 Capitol Records v. Vimeo, 826 F.3d 78 (2d Cir. 2016).

24 See https://rightsmanager.fb.com/.


26 https://creativecommons.org/.

27 https://creativecommons.org/share-your-work/licensing-considerations/.


Bibliography


Barlow (1994), ‘The Economy of Ideas: Selling Wine without Bottles’, 2.03 *Wired*


Byrne (2013), ‘The Internet will suck all creative content out of the world’, *The Guardian*

Dusollier, Ker, Iglesias and Smits (2014), *Contractual Arrangements Applicable to Creators: Law and Practice of Selected Member States*, Study, European Parliament, Committee on Legal Affairs


Ginsburg (2010), ‘The U.S. Experience with Mandatory Copyright Formalities: A Love/Hate Relationship’, 33 *Columbia Journal of Law and the Arts*, 311


IFPI & IMPALA (coordinators), *Securing a sustainable future for the European music sector*, Letter addressed to Jean-Claude Juncker on 29 of June 2016

Kantar Media (2016), *Online Copyright Infringement Tracker, Latest wave of research Mar 16 - May 16, Overview and key findings*, UK Intellectual Property Office


Mazziotti (2013), *Copyright in the Digital Single Market*, Centre for European Policy Studies (CEPS), Brussels


Quintais (2018), *Global Online Piracy Study Legal Background Report*, Information Law Institute, University of Amsterdam (IVIR Study)


Proposed copyright rules: bad for small publishers, European consumers and online services

Richard Gingras
VP, News

Published Dec 6, 2018

Copyright rules give news publishers rights over how their work is used. Europe is updating these rules for this digital age, and that's a move Google supports. But the European Parliament's version of a new copyright directive—specifically Article 11 and its recital 32—will have unintended consequences for smaller news publishers, limit innovation in journalism and reduce choice for European consumers. We urgently call on policymakers to fix this in the final text of the directive.

Let us be clear on one thing: Article 11 seeks to protect journalists and their work, and we agree with that goal. We care deeply about supporting the broader news industry because journalism is critical to the functioning of a free democracy. And we built Google to provide everyone with equal access to information.

However, Article 11 could change that principle and require online services to strike commercial deals with publishers to show hyperlinks and short snippets of news. This means that search engines, news aggregators, apps, and platforms would have to put commercial licences in place, and make decisions about which content to include on the basis of those licensing agreements and which to leave out.
Effectively, companies like Google will be put in the position of picking winners and losers. Online services, some of which generate no revenue (for instance, Google News) would have to make choices about which publishers they’d do deals with. Presently, more than 80,000 news publishers around the world can show up in Google News, but Article 11 would sharply reduce that number. And this is not just about Google, it’s unlikely any business will be able to license every single news publisher in the European Union, especially given the very broad definition being proposed.

This would mostly benefit larger players. One analysis has forecast that in Germany, small publishers would receive less than 1 percent of the revenue generated by a so-called ancillary copyright—whereas the largest publishing group alone would receive 64 percent. Smaller newsrooms and overall online news diversity will be impacted as a result.

Because so much of the conversation in Brussels is driven by larger publishing organizations, the small publishers who raise this concern are not heard. Why are large influential companies constraining how new and small publishers operate? Particularly at a time when news business models continue to evolve, new, small, and innovative publishers need flexibility. The proposed rules will undoubtedly hurt diversity of voices, with large publishers setting business models for the whole industry. This will not benefit all equally.

Not only might this harm individual news publishers, it also seriously risks reducing consumers’ ability to discover and access a diversity of views and opinions. Unlike people in other parts of the world, European citizens may no longer find the most relevant news across the web, but rather the news that online services have been able to commercially license. We believe the information we show should be based on quality, not on payment. And we believe it’s not in the interest of European citizens to change that.

Today we drive economic value to publishers by sending people to news sites over 10 billion times a month. That free traffic has enabled many smaller or emerging publishers to get discovered, grow a business, and find success online. A Deloitte study found that each user visit was worth on average between €0.04 and €0.08 to publishers. That means real business value to European publishers, every year.

We recognize the news industry is undergoing substantial change as publishers around the world transition to digital. We’ve been working with EU institutions to develop workable solutions that benefit journalists and publishers. We’ve invested in creating tools to help publishers increase subscription revenue and enable mobile sites to be much faster, so that they can grow their audiences and their revenue.
Thousands of news publishers use Google advertising services where they retain 70 percent and more of the revenue that's generated.

There is a way to avoid the unintended consequences of Article 11. The copyright directive should give all publishers the right to control their own business models and destiny by giving them the choice to waive the need for a commercial license for their content. Publishers—big and small—should continue to be able to make their own choices about how their content can be discovered and how they want to make money with that content. The exact language of the new rules is being determined in the next few weeks. Now is not the time to stifle innovation in news or limit access to quality journalism.
The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools

Evan Engstrom and Nick Feamster
March 2017
Table of Contents

Executive Summary ................................................................. i
I. Introduction ........................................................................... 1

II. Legal Background .............................................................. 3
   A. The Origins of the DMCA .................................................. 3
   B. Policy Rationale for Putting Burden of Policing on Owners .... 4
   C. Filtering Technologies Under the DMCA ............................ 5
   D. Cooperation Between OSPs and Copyright Owners .......... 8
   E. The Success of Congress's Voluntary Approach to Content Filtering ................................................. 9
   F. New Calls for Mandatory Filtering ................................... 10

III. Technical Analysis of Common Filtering Tools .............. 12
   A. Overview of Existing Filtering Mechanisms ...................... 12
      1. Content Metadata Search ............................................. 12
      2. Hash-Based Identification .......................................... 14
      3. Audio and Video Fingerprinting ................................. 15
   B. Case Study: Echoprint .................................................... 17
      1. Accuracy of Echoprint in Controlled Experiments .......... 19
      2. Echoprint's Limitations .............................................. 19
   C. Fundamental Problems with Filtering Technology ............. 20

IV. Negative Implications of Filtering Requirement ............... 25
   A. Filtering Tools Are Prohibitively Expensive for Many Small OSPs ......................................................... 26
   B. A Filtering Obligation Would Undermine the Certainty of the Safe Harbor ............................................. 28
   C. A Filtering Obligation Would Likely Have Minimal Impact on Infringement .......................................... 31

V. Conclusion ............................................................................ 32
Executive Summary

Nearly twenty years ago, Congress passed the Digital Millennium Copyright Act (DMCA), establishing copyright rules tailored for the emerging internet ecosystem. In order to provide legal certainty for internet startups and investors, as well as to encourage websites and copyright owners to cooperate in addressing online infringements, the DMCA created “safe harbors” for online service providers (OSPs) that limit exposure to secondary infringement claims for the actions of their users. At its core, the DMCA exempts OSPs from monetary liability for user infringements so long as the OSP removes access to infringing material when it becomes aware of it. But, because OSPs cannot readily identify potentially copyrighted material and determine whether or not its use is authorized, the DMCA places the burden of policing infringements squarely on copyright owners.

The DMCA’s limited liability regime prompted an unprecedented boom in internet activity to the benefit of users, OSPs, and creators alike. Although the DMCA has succeeded admirably in fostering the growth of the internet, some policymakers and copyright industry lobbyists have advocated for drastic changes to the copyright system to force OSPs to implement content filtering technologies in order to obtain the protections of the safe harbor and potentially punish them if and when those filters fail. Ultimately, these proposals hinge on a misunderstanding of the technical capabilities and likely effect of filtering technologies. In this paper we examine the functionality and inherent limitations of the most common filtering technologies to demonstrate why a mandatory filtering regime would pose grave dangers to the viability of the internet ecosystem in exchange for a minimal effects on online infringement:

- **Metadata Filtering**: Digital files are often labeled with information describing the contents of a given file (aka “metadata”). In the case of a song file, metadata could cover the song’s title, performer, and length. Automated programs can search a group of files for particular metadata matching a target copyrighted work and mark matching files for removal requests. These tools are inaccurate and easily circumvented, as a file’s metadata is often inaccurate and can be easily manipulated or encrypted to avoid detection. Similarly, metadata searches often misidentify non-infringing work, since two pieces of different content can share the same metadata (two songs with the same title, for example).

- **Hash-Based Filtering**: Using a file as the input in a specialized mathematical algorithm called a hash function produces a unique alphanumeric code (a “hash”) that can be used to identify specific files. Automated programs can compare the hash of a reference file containing a copyrighted work against a database of hashes for different files hosted on
an OSP to identify copies of the target work. But, because any alteration in a file’s data (such as changing the encoding format, shortening a song by a fraction of a second, or compressing the file) will produce an entirely different hash, hash-based filtering can only identify exact matches of specific files but cannot identify modified files containing the same underlying copyrighted work. Like metadata searches, hash-based filtering tools are also easy to circumvent through file manipulation and encryption.

- **Content Fingerprinting:** Unlike metadata search tools (which look at the metadata of particular files) and hash-based filtering tools (which turn the bits comprising a file into a unique identifier), fingerprinting tools examine characteristics of the underlying media content—such as the frequency values in a song file—to make identifications. While these tools are more robust to alterations in the file itself than hash or metadata-based techniques, they are similarly limited in their capacity to identify content. Fingerprinting tools are narrowly tailored to particular media types (an audio fingerprinting tool cannot be used to match copyrighted text files), and such tools only exist for a small subset of the many types of copyrighted content available online. Because fingerprinting tools require access to the underlying media content, they cannot be used to process encrypted files or torrent files, and they are not consistent with search platform functionality.

Critically, all content filtering technologies are at best capable of simply identifying the contents of a file, not making the often complex determination as to whether the use of a particular file constitutes an infringement. Furthermore, no filtering technology can process encrypted files.

For the great majority of OSPs that do not host large volumes of infringing content, the cost of filtering technologies far exceeds their benefit in limiting infringement. Such tools can require significant human and financial capital to maintain, and most of the websites trafficking in infringing material are based overseas or are otherwise beyond the reach of U.S. copyright law. Ultimately, conditioning access to the safe harbor on an OSP implementing a filtering technology would undermine the certainty that the DMCA has created without a commensurate impact on infringement. Policymakers should resist calls to upend the DMCA on the illusory hopes that filtering technologies can replace the commonsense copyright regime that has allowed the internet to flourish.
I. Introduction

In the span of a generation, the internet has gone from relative obscurity to near universality. Almost 90 percent of Americans directly use the internet, and essentially all aspects of the modern economy depend on internet access.¹ Given this fast pace of innovation, it is something of a surprise that many of the laws governing the internet that were created when the Web was emerging are still effective today.

One of the most critical of these laws is the Digital Millennium Copyright Act (DMCA). Passed with almost unanimous, bipartisan support in 1998, the DMCA provides rules clarifying the circumstances under which online companies may be held liable for the copyright infringements of their users, encouraging cooperation between websites and rightsholders to address online infringement and driving investment in the sector. Congress determined that these goals were best met by placing the burden of identifying infringing material on copyright holders, not online service providers (OSPs). At its core, the DMCA establishes that if a website removes access to infringing content when a copyright owner provides notice of the infringement, it cannot be held monetarily liable for the infringement.

The DMCA’s limitation on liability and division of responsibility between copyright owners and websites spurred the internet’s boundless growth. But, in recent years, copyright owners have argued that the internet’s rapid expansion has led to an equivalent increase in online infringement, such that the DMCA’s division of responsibilities between OSPs and copyright owners no longer makes sense. According to this view, websites should be required not only to remove access to a particular infringing file upon request but also to (1) proactively identify and delete other copies of the copyrighted work at issue and (2) guarantee that no future copies of that work ever appear their sites again. Arguments for this shift from a “notice-and-takedown” standard to “notice-and-staydown” depend on the notion that new technologies make it easy for OSPs to automatically and accurately

¹ Monica Anderson and Andrew Perrin, 13% of Americans Don’t Use The Internet. Who Are They?, PEW RESEARCH (Sept. 7, 2016), http://www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/
locate and remove infringing content. Proponents of “notice-and-staydown” are advocating for laws that require websites hosting user-posted content to implement such “filtering” technologies, rather than treating these automated technologies as voluntary measures, as they currently are under the DMCA.

Although filtering technologies have some role to play in the online ecosystem to identify and remove infringing material, their many inherent limitations make them incapable of fully addressing copyright infringement. For many websites, the costs and problems associated with filtering technologies far outweigh their benefits for reducing infringement. All of the filtering technologies available today, from hash-based filtering to fingerprinting, are limited in their capacity to accurately identify content. Simple systems such as hash-based filtering and metadata searches rely on crude approximations of the contents of a given file, making them easy to circumvent and prone to false identifications. More complex “fingerprinting” tools that examine the underlying contents of a file are not much more effective, as they also suffer from regular false positives and can at best be used to identify only a small subset of the many different types of copyrightable media that OSPs host. Although there are fingerprinting tools available to scan and compare audio, video, and image files, no such tools exist to process other forms of copyrightable content, such as software. And, since these filtering tools require access to the complete, raw, unencrypted content of files, they cannot process encrypted files or torrents. As a result, the range of infringing activity that filtering tools can effectively address is rather narrow. And, even for media types for which filtering tools exist, such tools are only capable of matching content, not determining whether the use of a particular work constitutes an infringement.

With increasing calls to modify the DMCA to require OSPs to proactively filter content, it is useful to consider how filtering tools actually work to shed light on whether they are likely to cause more harm than good for most OSPs, users, and creators. This paper examines the technical functioning of common filtering tools and their effects on infringement and the internet ecosystem to highlight how the DMCA’s approach to combatting online infringement is, despite its age, still well-suited to the technological realities of the internet and the goal of promoting the growth of both creative production and internet services at large.
II. Legal Background

A. The Origins of the DMCA

At the dawn of the internet era, it became clear that one of the internet’s central virtues—its capacity to facilitate the almost instantaneous distribution of information of all kinds, including copyrighted material, to any connected point on the globe—presented complications for existing copyright law. Congress recognized that the decentralized nature of the internet meant that it would be essentially impossible for OSPs to know about, much less review and control, all infringing activities occurring on their systems. Existing principles of copyright infringement liability—both direct and secondary—seemed an awkward fit for this new medium. Early decisions on copyright liability for OSPs applied different tests and came to often conflicting conclusions. Should OSPs be held directly liable for the infringing activities of their users? Does the operator of an intangible, unmoderated service like an online bulletin board exercise the same ability to control infringing activity as the owner of a dance hall or swap meet? Is the control an OSP exercises over its users more like the control a landlord exercises over her tenants or like a radio station over its disc jockeys? Congress understood that allowing courts to make these determinations on an ad hoc basis would preclude an OSP from knowing in advance whether or not it could be held liable for copyright infringement, severely discouraging investment in the sector.

To help mitigate this uncertainty while at the same time encouraging OSPs and rightsholders to cooperate in limiting online copyright infringement, Congress passed the DMCA, which at its core creates a legal “safe harbor” that allows compliant OSPs to avoid monetary liability for secondary copyright infringement.
claims.6 Essentially, as long as an OSP removes access to allegedly infringing material that it becomes aware of and terminates users who repeatedly infringe copyrights,7 it cannot be held monetarily liable for the copyright infringements of its users.

Since the obligations and protections of the DMCA’s safe harbor depend on whether or not an OSP has knowledge of specific infringing activity, defining the circumstances when an OSP should be held to have such knowledge is central to the operation of the statute. Critically, Congress determined that an OSP should not be assumed to have knowledge of all user activities and established that eligibility for the safe harbor does not require an OSP to “monitor[] its service or affirmatively seek[] facts indicating infringing activity.”8 Rather, the “DMCA notification procedures place the burden of policing copyright infringement—identifying the potentially infringing material and adequately documenting infringement—squarely on the owners of the copyright.”9 As such, to qualify for the safe harbor, an OSP must prevent access to infringing material if it independently becomes aware of such material or if a copyright owner or its agent sends a statutorily compliant notice of claimed infringement to the OSP, but it need not take proactive steps to seek out infringing material on its site.

B. Policy Rationale for Putting Burden of Policing on Owners

Congress decided to refrain from automatically imputing knowledge of user infringements on OSPs and to place the burden of policing infringement on copyright owners in large part because “[c]opyright holders know precisely what materials they own, and are thus better able to efficiently identify infringing copies than service providers…who cannot readily ascertain what material is copyrighted and what is not.”10 Given the informational asymmetry between OSPs and copyright owners, requiring OSPs to identify potential infringements would be

6 See e.g 144 Cong. Rec. S11,889 (daily ed. Oct. 2, 1998) (statement of Sen. Hatch) (DMCA meant to provide OSPs with “more certainty … in order to attract the substantial investments necessary to continue the expansion and upgrading of the Internet.”).
7 An OSP that wants to obtain the protection of the DMCA’s safe harbor must also meet several other requirements, such as designating an agent to receive takedown notices. 17 U.S.C. § 512.
8 17 U.S.C. §12(m).
9 Perfect 10, Inc. v. CCBill LLC, 488 F.3d 1102, 1113 (9th Cir. 2007).
10 UMG Recordings, Inc. v. Shelter Capital Partners LLC (UMG II), 718 F.3d 1006, 1021-22 (9th Cir. 2013).
inefficient, impractical, and prohibitively expensive for many OSPs. As the Senate noted in its report accompanying the DMCA, OSPs—often fledgling startups—should not be expected “to determine whether [a] photograph was still protected by copyright or was in the public domain; if the photograph was still protected by copyright, whether the use was licensed; and if the use was not licensed, whether it was permitted under the fair use doctrine.”\textsuperscript{11} It is difficult to imagine any small OSP being able to thoroughly monitor all of the content on its site and make accurate infringement determinations on a case-by-case basis.

Because most startups are ill-equipped to perform this level of analysis and in any event typically do not have enough infringement occurring on their platforms to warrant the imposition of a stringent content monitoring obligation, Congress recognized that placing the burden of policing infringement on OSPs would simply result in some combination of severely curtailed investment in online startups and the systematic removal of non-infringing material. Because copyright holders can seek statutory damages awards well in excess of the actual harm caused by an act of infringement,\textsuperscript{12} OSPs already have strong incentives to process takedown notices for material that is arguably non-infringing, as the cost of a statutory damages award usually far exceeds the benefit of resisting a questionable removal request. Had Congress determined that OSPs were obliged to proactively monitor for potentially infringing material, and that they could be held liable for infringements for which they had no knowledge, it is unlikely that any Web 2.0 companies would have found investors in light of the constant threat of unforeseen copyright liability. As a result, the many OSPs that allow users to distribute, share, and store content online would not exist today.

\textbf{C. Filtering Technologies Under the DMCA}

Underlying Congress’s decision to relieve OSPs from the burden of monitoring user activities and searching for infringing material was the recognition that such a task was technologically infeasible. As then-Senator John Ashcroft noted in the lead-up to the DMCA’s passage:

\begin{itemize}
\item \textsuperscript{11} S. Rep. No. 105–190, at 48.
\item \textsuperscript{12} 17 U.S.C. § 504(c).
\end{itemize}
It would be impossible for any carrier to review all of the material; and we cannot create a legal obligation that is technologically impossible to satisfy. Clearly, the potential for copyright infringement is real—as real as the impossibility of requiring a service provider to monitor every communication, including e-mail, homepages, and chat rooms [for infringing activity].

Although Congress thoroughly considered how best to encourage the cooperative development and deployment of technologies that would help minimize copyright infringement on the internet, it ultimately decided to exempt OSPs from an obligation to affirmatively monitor their services for possible infringements, whether through manual review or automated technologies.

As Congress anticipated, despite the lack of an affirmative mandate to deploy technologies to identify potentially infringing content, many such technologies emerged in the years after the DMCA’s passage, and courts were tasked with determining how the use of such programs might affect an OSP’s obligations under the DMCA. Perhaps the most direct analysis of the interplay between filtering technologies and the DMCA’s requirements for OSPs came from the Central District of California in *UMG Recordings, Inc. v. Veoh Networks, Inc.* In *Veoh*, the court granted summary judgment against UMG on its secondary infringement claims against Veoh, a user-generated content (UGC) video platform, holding, *inter alia*, that Veoh’s decision to adopt two different filtering technologies did not affect its ability to claim the protections of the DMCA’s safe harbor. UMG argued that even though Veoh implemented a hash-based filtering system early in its existence, its delay in adopting an allegedly more robust filtering system amounted to willful blindness of the infringing activity occurring on its system. UMG claimed that this delay should deprive Veoh of the safe harbor, as it was intentionally avoiding gaining knowledge of infringements. The court rejected this argument on the grounds that the DMCA does not impose “an obligation on a service provider to implement filtering technology at all, let alone technology from the copyright

---

16 See section III.A.2., infra.
17 UMG (I), 665 F. Supp. 2d at 1111.
holder’s preferred vendor or on the copyright holder’s desired timeline.”18 Rather, the court noted that Veoh’s voluntary adoption of filtering technologies demonstrated its good faith efforts to avoid or limit storage of infringing content.19 Similarly, the court held that “Veoh’s ‘right and ability’ to implement filtering software, standing alone or even along with Veoh’s ability to control users' access” did not mean that Veoh had the “right and ability to control [infringing] activity” on the site such that it should be excluded from the safe harbor pursuant to 17 U.S.C. 512(c)(1)(B).20

If courts were to find that the availability of superior filtering systems or the ability to search for potentially infringing files establishes – without more – that a service provider has “the right and ability to control" infringement, that would effectively require service providers to adopt specific filtering technology and perform regular searches. That, in turn, would impermissibly condition the applicability of section 512(c) on “a service provider monitoring its service or affirmatively seeking facts indicating infringing activity.”21

As other courts have noted,22 conditioning eligibility for the safe harbor on an OSP’s decision to implement content filtering tools would directly violate Congress’s intent:

[The DMCA] is not intended to discourage the service provider from monitoring its service for infringing material. Courts should not conclude that the service provider loses eligibility for limitations on liability under section 512 solely because it engaged in a monitoring program.23

Although it is well-established that the DMCA neither requires an OSP to implement a filtering system to obtain the protection of the safe harbor nor punishes an OSP that voluntarily adopts such a system by imparting on the OSP disqualifying knowledge of infringing activity, an OSP’s decision to implement or

---

18 Id.
19 Id. at 1112.
20 Id. at 1113.
21 Id.
forego filtering technologies has on occasion impacted a court’s consideration of an OSP’s intent for purposes of a secondary infringement claim. Though it was not strictly relevant to the outcome in *UMG Recordings, Inc. v. Veoh Networks, Inc.*, the court noted that Veoh’s adoption of two different filtering systems was evidence of Veoh’s “good faith efforts to avoid or limit storage of infringing content.”24 In *Metro-Goldwyn-Mayer, Inc. v. Grokster*, the Supreme Court cited the defendant OSP’s failure to “develop filtering tools or other mechanisms to diminish the infringing activity using their software” as additional evidence of their “intentional facilitation of their users infringement,” though it cabined this dicta by noting that “in the absence of other evidence of intent, a court would be unable to find contributory infringement liability merely based on a failure to take affirmative steps to prevent infringement.”25

D. Cooperation Between OSPs and Copyright Owners

Though there has been relatively little judicial analysis of the role that filtering technologies play under the DMCA, the congressional record, statutory text, and legal interpretations all clearly establish that the use of filtering technologies is purely voluntary for OSPs and does not impose any barrier to accessing the safe harbor.26 Despite some concerns at the time of the DMCA’s passage that making the use of filtering technologies optional would not adequately incentivize OSPs to develop and adopt tools to address online infringements,27 many such tools have been developed and implemented on web platforms, at least on larger OSPs where the greater volume of user activity makes these tools more useful. For all but the largest OSPs, the small scope of infringing material on their portals renders

---

24 UMG (I), 665 F. Supp. 2d at 1112.
25 Metro-Goldwyn-Mayer, Inc. v. Grokster, 545 US 913, 939; but see Capitol Records, LLC v. Vimeo, LLC, 972 F. Supp. 2d 500, 534 (S.D.N.Y. 2013) (“Plaintiffs further contend that evidence of inducement may be found in Vimeo’s failure to implement filtering technologies that could be used to locate infringing content. But just because Vimeo can exercise control does not mean that it must.”).
26 This rule is constrained in theory by language in the statute establishing that an OSP seeking the safe harbor may be required to proactively monitor for user infringements “to the extent consistent with a standard technical measure.” See 17 USC §512(m). This limitation is only theoretical, since there are no such “standard technical measures” in existence. The statutory definition of “standard technical measures”—limited to technologies “used by copyright owners to identify or protect copyrighted works” that have been “developed pursuant to a broad consensus of copyright owners and service providers in an open, fair, voluntary, multi-industry standards process” and “do not impose substantial costs on service providers”—is very narrow and appears incompatible with all past and present filtering technologies. 17 U.S.C. § 512(i)(2) (emphasis added). See also L. Gallo, The Impossibility of “Standard Technical Measures” for UGC Websites, 34 COLUM. J.L. & ARTS 283 (2011). Considering the severe negative impacts that a mandatory content policing regime would have on the online ecosystem, it is no surprise that Congress limited “standard technical measures” to those technologies that are so uncontroversial as to have near unanimous approval.
27 See Mehra & Trimble, supra note 14, at 690-691.
filtering tools impractical and unnecessary.\textsuperscript{28} As Congress hoped, content owners and OSPs have entered into a variety of cooperative arrangements to help mitigate online infringement (and even allow copyright owners to monetize infringing user uploads rather than simply removing them, as with YouTube’s Content ID). While early examples of such cooperation (such as the Principles for User Generated Content Services (“UGC Principles”) from 2007\textsuperscript{29}) were predicated in part by OSPs’ concerns regarding potential liability, many of the arrangements in place today often derive from cooperative business partnerships between OSPs and copyright owners.

E. The Success of Congress’s Voluntary Approach to Content Filtering

Congress’s decision to encourage voluntary cooperation between content owners and OSPs to combat online infringement rather than placing the burden of searching out infringements on OSPs has succeeded in diminishing legal uncertainty regarding OSP liability for user infringements, allowing rightsholders to efficiently address copyright infringement on the many OSP platforms that do not experience significant infringing user activity, and fostering collaboration between OSPs and rightsholders on sites with enough user traffic such that more proactive measures are sensible. More critically, it has furthered the fundamental purpose of copyright law by facilitating the growth of a vibrant ecosystem of online portals that have helped spur a boom in creative production since the DMCA’s inception. The economic value of the worldwide entertainment industry grew almost 66 percent from 1998 to 2010, and U.S. households spent almost 15 percent more on entertainment as a percentage of income in 2008 than in 2000.\textsuperscript{30} In terms of actual content creation, the years since the DMCA’s passage have been amongst the most prolific in history.\textsuperscript{31} OSPs have seen similar growth, as the internet sector

\textsuperscript{28} Gallo, supra note 26, at 295 (“[A]ll major right holders and all major UGCs...employ some form of the fingerprinting technology.”)

\textsuperscript{29} The UGC Principles are a voluntary, non-binding agreement between content owners and OSPs providing that content owners will not sue signatory OSPs that, among other things, implement “effective content identification technology.” \textit{Principles for User Generated Content Services}, \url{http://www.ugcprinciples.com} (last visited March 20, 2017).


\textsuperscript{31} Between 2001 and 2010, the number of songs identified in the Gracenote database jumped from 11 million to over 100 million; between 2002 and 2010, the number of new books available on the market increased more than twelvefold. See Masnick supra note 30, at 3.
contributed nearly $1 trillion to the U.S. economy—nearly 6 percent of real GDP—in 2014.\(^\text{32}\)

**F. New Calls for Mandatory Filtering**

Despite the success of the DMCA, some copyright holders and policymakers have pushed to rewrite the law, arguing that the internet has grown in a way that was not anticipated by the DMCA’s drafters such that its provisions no longer make sense in the current online environment.\(^\text{33}\) They claim that the incredible growth of the internet and the exponential increase in content available online renders the DMCA’s “notice and takedown” process ineffective at addressing online infringements, as content owners cannot identify and request removal of infringing content as quickly as new infringing content is posted.\(^\text{34}\) Content owners complain that because the DMCA only requires OSPs to disable access to specifically identified infringing files and imposes no obligation to independently locate and remove other copies of works identified in takedown requests, they must send a takedown notice for each copy of a particular work on any given website—a Sisyphean task, according to copyright holders.\(^\text{35}\)

To address the DMCA’s perceived shortcomings, many content owners have argued that OSPs should be obligated to implement content filtering technologies to pre-screen all user uploads as a requirement to qualify for the safe harbor.\(^\text{36}\) Under this proposal, “websites will be immune from copyright liability—both direct and indirect—as long as they can show that they employed the best filtering technology available on the market at the time the alleged infringement occurred.”\(^\text{37}\) These proposals are often conjoined with so-called “notice-and-staydown” obligations under which a standard takedown notice from a copyright


\(^{33}\) See, e.g., Comments of Universal Music Group, U.S. Copyright Office, Section 512 Study, at 9 (2016) (“Whatever its original intent, Congress simply could not have contemplated either the massive scope or the diverse methods of online infringement when it passed Section 512 in 1998”).


\(^{35}\) See Comments of the Motion Picture Association of America, U.S. Copyright Office, Section 512 Study, at 5 (2016) (“Because infringing content is uploaded on a mass scale, requiring copyright owners to provide individualized URLs in order to trigger service providers’ obligation to act further ensures that the process will be repetitive, mechanical and ineffective.”).

\(^{36}\) See *UMG Comments*, supra note 32 (“UMG respectfully submits that a service provider wishing to take advantage of the Section 512 safe harbors must, at a minimum, implement effective content identification technologies.”).

owner for a particular infringing file would trigger an obligation for the OSP to locate and remove every other instance of the identified work from the website and prevent future uploads of that work. Since manually comparing all uploads against a database of prior takedown requests is logistically impractical, notice-and-staydown proposals tacitly include a mandatory content filtering obligation and would punish OSPs when those filtering tools fail to remove all infringing content.

Although the DMCA’s voluntary approach to OSP content filtering remains in place domestically, policymakers in Europe recently proposed an update to the European Commission’s copyright regime that includes a requirement for OSPs to implement content filtering technologies. The European Commission’s (EC) proposal is a significant departure from existing EU copyright law, which largely mirrored the DMCA’s notice-and-takedown approach. A leaked document outlining the EC’s policy rationale for proposing a mandatory filtering obligation reflects a misunderstanding of the technological and economic realities of content filtering, implying that filtering technologies are capable of accurately identifying a wide range of content and are available to startups for negligible costs.

Proposals from the EC and copyright industries that would require OSPs to implement filtering technology misunderstand the inherent technical limitations of these technologies, as well as the costs and consequences of mandating their use. While using certain filtering tools may be an effective way for some OSPs to address online infringements, their cost and limited utility militate against mandating universal adoption or conditioning qualification for safe harbor protection on their use. Given the technical shortcomings of content filtering tools, requiring OSPs to implement such technologies will likely have minimal effect on copyright infringement but seriously harm the OSPs that have helped make the internet the dominant cultural and economic medium of our time.

III. Technical Analysis of Common Filtering Tools

As with all areas of computing, there have been significant advances in content filtering technologies since the adoption of the DMCA. The technologies available today can more quickly scan files to identify the underlying content with more adaptability to potential alterations in the file format or encoded media. However, all of the tools currently available to locate potentially infringing material are subject to severe limitations with respect to their accuracy and adaptability. Additionally, given the wide range of copyrightable content, there are many types of media content for which no filtering tools currently exist. To help explain why filtering technologies cannot realistically be expected to accurately identify all infringing content or otherwise eliminate online copyright infringement, it is helpful to examine the basic functionality of the most commonly employed filtering technologies and to describe their uses and limitations.

A. Overview of Existing Filtering Mechanisms

1. Content Metadata Search

One common technique for identifying content online, in use since at least the early 1990s with the advent of Gopher, involves searching files—either manually or through the use of automated programs—according to data that accompany the actual content, or “metadata.” The term “metadata” refers to the fact that the content is annotated with additional data as opposed to simply being described by the content itself; in other words, it is structured information about the associated media resource. Metadata may contain various information about a given file; in the case of an audio or video media file, for example, metadata may include the content’s title, data, file size, length, encoding rate, and so forth. Metadata can take on many formats, but it is generally structured in a way that makes it easy to search. Given such data about associated media, a simple takedown strategy entails searching an OSP for offending content based on its metadata and removing any matches. For example, an automated script could search for content...

based on metadata associated with the content, such as the document’s title (e.g., a song title), author, or artist. In the case of YouTube, for example, the metadata include information such as the content’s author (or owner), a title, a text description, a set of text-based keywords, the time at which the video was uploaded, the duration of the file, the name of the raw file, and various viewing statistics (e.g., the number of times the video has been viewed). In principle, one could perform a search on any of this metadata to retrieve files with associated matching metadata. Note that it is possible to fetch this metadata programmatically (e.g., with an automated script) without ever examining the contents of the file itself. Similar metadata exist for content that is hosted on other OSPs. Searching for content based on its metadata is simple and efficient, as it does not involve direct analysis of the content itself.

Metadata searches are efficient in that they allow for the quick analysis of a large volume of files without needing to actually download any files and can be used across a variety of media.\(^\text{43}\) However, they are imprecise and often inaccurate. Clearly, metadata searches will be unable to identify content if the metadata of a particular file do not accurately describe its content. Content is often mislabeled, either accidentally or intentionally. Similarly, two different pieces of media content can have the same metadata—a book and a movie with the same title, for example.\(^\text{44}\) As a result, a search that aims to identify offending content for removal simply based on metadata may be subject to both false positives (in cases where portions of the metadata, such as a title, match the title of infringing content, but the content itself does not match the infringing content) and false negatives (in cases where content that corresponds to infringing content is labeled with metadata that do not match the metadata for the content). For example, audio, video, or other content might be re-titled, re-encoded, shortened, or otherwise modified in ways that affect the content metadata and thus the ability to perform a metadata search to identify corresponding infringing content. In particular, converting a media file from one format to another can often alter or eradicate

\(^{43}\) As described in more detail infra, some fingerprinting tools are limited to a particular media type. A tool designed to identify the contents of video files will not work on text, for example. And no such fingerprinting tools even exist for many types of copyrightable media.

\(^{44}\) See, e.g., Ernesto, “Warner Bros. Admits Sending Hotfile False Takedown Notices,” TORRENTFREAK, (Nov. 10, 2009) https://torrentfreak.com/warner-bros-admits-sending-hotfile-false-takedown-requests-111109/ (“Hotfile pointed out that [Warner’s] automated process resulted in the removal of many files that do not belong to Warner. The movie studio admits this and confirms that while searching for ‘The Box (2009)’ many unrelated titles were removed. ‘Warner admits that its records indicate that URLs containing the phrases ‘The Box That Changed Britain’ and ‘Cancer Step Outsider [sic] of the Box’ were requested for takedown through use of the SRA tool.’”)
metadata, making searches for metadata inaccurate or otherwise impractical. In short, the accuracy of metadata searches is limited by the accuracy and precision of the metadata associated with every file at issue. Because it is easy to alter the metadata of a file, and because a file’s metadata are not necessarily unique markers for particular content, metadata searches are not reliable for accurately identifying potentially infringing content.

2. Hash-Based Identification

Another approach to automated identification of content involves representing a piece of content (e.g., an audio or video file) by a content hash. A hash is a numerical representation of a file that is significantly smaller than the original file. The hash value of a file is computed using a cryptographic hash function that takes the file as input. A file’s hash nonetheless uniquely represents the contents of the file, so that files with different contents will generally have different associated hash values. Commonly used cryptographic hash functions have a property called collision-resistance that aims to guarantee that two distinct inputs (i.e., files) will not produce the same hash value. Moreover, even slight alterations to a file will generally result in completely different hash values. As such, except in very rare cases of hash collisions, hashes are effectively perfect identifiers of a particular file.

Additionally, because the hash of a file is generally much smaller than the file itself (a hash might be, for example, 256 bits—a small fraction of the overall size of the file, since video files can easily exceed 100 megabytes), it is much more convenient to analyze hashes than the full files themselves. This smaller size lends itself to convenience for automated search and identification as well, since it is much easier to maintain a reference database of hashes than full files. Rather than relying on the accuracy of the metadata or requiring close examination of the entire file, an automated takedown mechanism could compute the hash value of a piece of content and compare that hash value against a database of hash functions that correspond to copyrighted content. If the same mechanism were required to compare full files against a database of full files, the process would require considerably more time, storage, and computing power. Consider, for example, that an an audio file that is encoded at 128 kilobits per second requires approximately one megabyte of storage per minute of audio; therefore, a typical audio file will require approximately 5 megabytes of storage—more than 100 times the size of
the corresponding hash value. In addition to the storage costs associated with storing and comparing files, the computational cost would be similarly higher, as comparing the contents of a larger file requires proportionally more time to read and process the file.

As with searching for metadata, searching for infringing content by hash values is more efficient than techniques that require analyzing full files; yet, a simple hash-based comparison also has multiple drawbacks. Notably, altering the original file in any way—shortening or excerpting it, re-encoding it, and so forth—will invariably alter the hash value. For example, if the same song is stored in two different file formats, one as a WAV and another as an MP3, each file will have a unique hash and the database of hash values against which a search is referenced will need to contain the hash values of both files to accurately identify both of them. A search for the hash of the MP3 version will not match the WAV file, even though the original data in each file were the same. As a result, just as altering a file’s metadata will undermine the accuracy of metadata-based content filtering techniques, infringing content may escape automated detection simply by virtue of being a slightly altered version of the original content.

3. Audio and Video Fingerprinting

Newer, more sophisticated tools for content identification, such as Audible Magic, PhotoDNS, Zefr, and Echoprint, involve a technique known as fingerprinting. Unlike hash-based filtering, these programs typically do not examine the bits in the media file (which may change with alteration and editing) to make identifications, but rather characteristics of the media itself—the notes in a song, for example, rather than the bits in the file that encode the song. By looking at the characteristics of the media itself rather than the bits comprising the entire file (as with hashing), a fingerprint computes a representation of the content that is more robust, even under various previously described alterations, edits, and modifications that would change the raw content of the media. More specifically, each note in a song could be represented by the presence or absence of specific frequency values; the volume of a particular note is roughly represented by some amplitude at the frequency corresponding to that musical note. Specific fingerprints may include information such as the amplitude of different frequency values over a sequence of time intervals. For example, an audio file may have more “energy” in higher or
lower frequencies (a high-pitched tone has more energy in higher frequencies). The amount of energy in each part of the frequency spectrum will differ across files, and that representation may be both unique to the file and robust to various transformations of the original files. One piece of content can often have multiple fingerprints: algorithms can produce fingerprints based on different excerpts of a file, for example, and different algorithms for producing fingerprints will naturally produce different fingerprints.

Because fingerprinting technologies rely on algorithms that process the underlying media content of a given file, they are naturally constrained to a small subset of copyrightable content. For example, because an audio fingerprinting tool’s central algorithm examines, say, the frequency values in a song file, it cannot be used to identify copyrighted photographs or software programs, which contain no audio frequency values. As such, to filter all files on a given site, an OSP would need to obtain a different fingerprinting tool for each type of media that is (or could be) hosted across the entire site. Considering the incredibly wide scope of copyrightable content, there are many types of content for which no fingerprinting tool exists, such as architectural designs or handmade items sold on Etsy.

For types of media that are amenable to fingerprinting, there are various algorithms that can produce such fingerprints; early versions of the tools to generate fingerprints for audio files, for example, generated fingerprints using the chroma vectors of successive smaller segments of the original file. In short, chroma vectors measure the relative intensities of the twelve pitches in the music scale within a given time interval (for example, how often the tone of “C#” is heard during a time interval and at what intensity). These vectors tend to be robust to noise such as ambient sounds, and they are also independent of volume and dynamics. Although relying on chroma vectors to produce an audio file fingerprint represents an improvement over previous hash-based matching mechanisms, the original fingerprinting techniques were not robust to distortions that may be introduced in over-the-air recordings; subsequent improvements to the original algorithm have enabled the encodings to be robust to these distortions.

---

Although audio fingerprinting tools overcome some of the limitations of tools based on hash or metadata searches—such as an improved capacity to accurately identify files that may have been altered in simple ways—these techniques remain at best an imperfect mechanism for automatically identifying copyrighted content. For example, fingerprinting techniques can be susceptible to false positives or false negatives when scanning files for which the encoded content has been altered in some way, and they are usually narrowly tailored to operate only on particular types of media; for example, a fingerprinting tool created to identify audio files cannot be used to identify software programs.

Most fingerprinting tools are proprietary, making it difficult to evaluate their technical functionality. Nevertheless, an examination of the mechanisms that underpin an open-source audio fingerprinting tool, Echoprint, is a representative illustration of how other fingerprinting programs operate. Although many such tools exist, they operate similarly to Echoprint and are thus subject to similar limitations.

**B. Case Study: Echoprint**

Echoprint is an open-source audio fingerprinting tool that is used by Spotify, among others. Because the tool is open-source, it is possible to analyze the tool in more detail than other contemporary fingerprinting tools, but other programs most likely rely on similar mechanisms. The main fingerprinting mechanisms described in this section are viewable in the Echoprint source code on GitHub. The designers of Echoprint have also released a white paper that describes the fingerprinting algorithm in detail, and many of the techniques are also summarized in a patent.

Echoprint has several components: (1) a tool to generate the “codes” or fingerprints for a particular media file; (2) a query server that stores the database of codes to match against; and (3) the codes themselves that are used to match against the codes that are generated from any particular media file. Given any media file or excerpt, the code generator can produce a code that is subsequently matched against codes stored at the query server.

---

The process Echoprint uses to generate a fingerprint relies on the relative timing of successive onsets (i.e., beat-like events) in the audio sample. Recently, these fingerprinting algorithms have also added “whitening” filters to remove resonance (i.e., echo) that may result from various artifacts that live recordings or over-the-air broadcasts may introduce. Although such a whitening filter makes fingerprinting more effective for accurately matching both live and over-the-air recordings with an original recording, this filtering process does not necessarily correct for more “natural” distortions that may result from a live performance (e.g., changes in beat onsets that may simply be an aspect of the performance itself, as opposed to signal distortion).

After applying a sequence of filters to the audio signal, the codes that Echoprint generates are based on analysis from eight separate frequency bands. One configuration uses frequency bands spanning from 0 to 5,500 Hz (i.e., samples per second), but other frequency bands may be used; the method does not depend on a precise specification of frequency bands. To identify matching content, the fingerprinting mechanism compares the relative timing between beat onsets in the audio that are present in each of these different frequency bands. An “onset” is the period of increased magnitude of the music sample, such as the start of a musical note or a beat; given onsets in each frequency band, the algorithm may associate a timestamp within each onset. The resulting fingerprint or “code” is based on the relative location of onsets within the audio sample.

Given this fingerprint, the algorithm identifies matching documents by comparing the fingerprint against a database of known tracks. During the fingerprinting process, each audio file is split into 60-second segments, with each adjacent segment overlapping by 30 seconds to produce more matching query segments for an audio file of a given length. The codes for any 60-second segment are represented as terms in an inverted index. An inverted index takes an audio sample and returns the corresponding document that contains that sample, as opposed to a conventional index, which typically takes a track or document ID and returns the corresponding file. The concept is analogous to a search index, whereby sets of keywords map to corresponding documents that contain those keywords. The track index and the segment number is the document ID; the underlying database facilitates a fast lookup of the document ID, given a code that is being
queried. Any 30-second query contains about 800 samples that are used to perform the inverted lookup, and the query server returns the documents with the most matches for each code term in the query. In practice, many documents may contain a set of matching hash values; each document has a corresponding score based on the number of matching hash keys across the segments in the file. The algorithm only returns a matching document (i.e., media file) if the score of a single document is significantly higher than all other candidates. Otherwise, the algorithm does not return a match.

1. Accuracy of Echoprint in Controlled Experiments

Echoprint’s designers performed an evaluation based on a database containing 30 million audio files. Depending on the length of the file and the type of encoding, the actual error rate for the algorithm varies between 1–2 percent, where an error could either be a false positive (the wrong file was identified), a false negative (the file was contained in the database but not identified), or a false accept (the file was not contained in the database at all, but a false match was detected). Given the reported error rates, one could thus expect the state of the art fingerprinting algorithm to misidentify about one or two in every 100 pieces of audio content. As a point of comparison for acceptable false positive rates, consider that mail service providers consider any false positive rate higher than about 0.1 percent too high to be used in practice for spam filters, due to the potential limitations on speech that could arise as a result of legitimate email messages being misidentified as spam. Accordingly, a 1–2 percent false positive rate for an automated filtering procedure is problematic for the same reasons, as such a technique would result in filtering legitimate content at rates that would frequently obstruct speech.

2. Echoprint’s Limitations

Even the most sophisticated audio fingerprinting technologies such as Echoprint have several fundamental limitations. Because Echoprint is based on the inter-onset intervals in the audio file, any distortion to these intervals will disrupt Echoprint’s ability to identify the corresponding audio file. Such distortions might be

commonplace in live performances of copyrighted content, but other transformations may also distort these intervals. One transformation that could distort these intervals might be excerpting certain tracks of a recording, or altering the playback rate of the audio.

Also, Echoprint, like other fingerprinting tools, is designed to operate on a single media type: audio. As such, it is limited in its efficacy, as an OSP cannot use Echoprint or other audio filtering technologies to accurately identify video, text, or binary executable files like software programs. Fingerprinting a binary executable, in particular, is not currently feasible, because binaries are not amenable to the signal processing techniques that Echoprint and other media fingerprinting technologies rely on. Any OSP that allows users to upload a variety of different media would likely have to implement a different fingerprinting tool for each different type of content; even then, it is unlikely that the OSP would be able to filter all of the content on the site.

In addition to its inability to identify files with audio distortions and its applicability to audio files only, Echoprint, like other fingerprinting tools, is subject to many other fundamental limitations that any filtering technology faces.

C. Fundamental Problems with Filtering Technology

Filtering technologies such as those described above have a variety of limitations, many of which make it fundamentally difficult to perform automated identification of content. We briefly outline these limitations below and explain why relying on automated filtering technologies for copyright enforcement may result in false positives and other unintended outcomes.

First, because all filtering technologies rely on examining and analyzing aspects of target files (whether the bits in a file as with hash-based filtering, identifying information describing a file as with metadata searches, or unique aspects of the media content of a file as with fingerprinting), all filtering tools can be evaded through basic manipulation of the file. In the case of identification based on metadata, even simple manipulation of the metadata (such as altering the title of the content) is sufficient to disrupt automated filtering. In the case of file content
hashes, the value of the resulting hash is based on a computation over the contents of the file; thus, even small changes to a file’s contents will result in a completely different hash computation. Thus, filters that rely on matches against hash values that correspond to infringing content can be undermined simply from small manipulations (such as shortening an audio file by a fraction of a second). Transformations and encryption will also undermine more sophisticated fingerprinting technology. Fingerprinting mechanisms that are based on aspects of the file’s contents (e.g., chroma vectors in the case of audio fingerprinting and techniques such as those used by Echoprint) are typically more resilient to minor manipulation because these types of features are preserved. Encryption, however, stymies even these more sophisticated fingerprinting techniques. When a file is encrypted, the entire contents of the message are completely obscured, making it impossible to analyze the contents of the file for specific patterns. Although it is possible to encrypt a file’s contents without encrypting the metadata, such a method would be subject to the previously discussed limitations of metadata-based fingerprinting—specifically, the file’s contents may not match the associated metadata, and a file could thus be identified based on incorrect metadata. If the file’s contents were encrypted, it would also be difficult to determine whether the contents matched the metadata, since the content itself could not be analyzed or inspected. Because all of these alteration techniques aimed at evading filtering technologies are easy to perform and in most cases do not require changes to a file that would render the encoded media less functional, filtering technologies are unlikely to be effective for mitigating copyright infringement.

Second, even if it were possible to consistently identify content accurately with filtering technologies, such technologies are not sufficient to consistently identify infringements with accuracy, as they can only indicate whether a file’s contents match protected content, not whether a particular use of an identified file is an infringement in light of the context within which the media was being used. It is often permissible to excerpt or otherwise refer to copyrighted content in contexts that are permitted by fair use (e.g., for educational purposes, in art). Although an automated algorithm could determine whether the content (or excerpt) matched known copyrighted content, such an algorithm would not be able to determine whether the particular use of a given file is infringing or not. Since making nuanced determinations as to the legality of a file’s use is appropriately left to legal
practitioners and juries, automated filtering technologies will never be sufficient to accurately police for infringing content.

Third, filtering technologies are only effective if an OSP has access to the file in question, making it difficult or impossible to use filtering tools on search platforms. Unlike OSPs that store material at the direction of a user (such as UGC sites), search engine platforms link to content that is hosted elsewhere often without downloading or otherwise processing the linked content. Although search engines build indexes based on content that they “crawl” from the web, ultimately these search engines map keywords and search phrases to links that are elsewhere on the internet, making it difficult, if not impossible, to control links to infringing content if the contents of a link were to change after the index was built. Similarly, search platforms cannot employ tools like hash-based filtering or fingerprinting, as they do not generally have access to linked files as would be necessary to compute a file’s hash value or fingerprint. If search platforms were required to actually download linked content and run the associated files through filtering technologies, the cost of indexing the content on the web would increase significantly, potentially crippling the efficiency of online search.

In some cases, it may be extremely difficult for the search engine to obtain the content in the first place. For example, protocols that facilitate file sharing such as BitTorrent rely on application-specific clients to locate, download, and assemble file segments (i.e., “chunks” in BitTorrent parlance) based on information contained in an index file (in BitTorrent, a so-called “torrent” file). Although the torrent file may be accessible via the web (and thus indexed by the search engine), the contents of the file that the torrent file refers to are impossible to retrieve without participating in the file-sharing protocol. As a result, search engines typically do not ever retrieve or store the contents of files shared using file-sharing software such as BitTorrent, even though the search engine may index the torrent files. In other cases where media might be streamed, the search engine would be unable to download the media in the first place, requiring any fingerprinting process to operate in real-time over a streaming audio file, which is impractical.

Similarly, the contents of any individual webpage are dynamic (i.e., generated programmatically with scripting languages such as Javascript), and are often even

---

personalized based on the user or device that is retrieving the content. Website owners have at times attempted to conceal copyright infringing activity, malware hosting, or other malfeasance by exploiting such dynamic behavior to conceal the infringing or offending content from the search engine itself. Ultimately, these limitations make filtering technologies ineffective for use in linking or other search-based applications.

Fourth, all automated filtering technologies depend on having a reference database for copyrighted content (such as matching fingerprints or hashes), but such a database is only possible if copyright owners submit accurate fingerprints hashes of content to the database in the first place. Reported problems with the accuracy of takedown notices raise questions as to whether the databases on which filtering tools depend are or will soon become suffused with false identifications where the reference hash or fingerprint does not actually correspond to the content it purportedly identifies. One study found that almost 5 percent of takedown notices in the chosen sample did not match the infringed work. In the context of the study, this amounted to around 4.5 million takedown requests. Similarly, many copyright owners simply refuse to provide any fingerprint data to some OSPs, rendering futile their efforts to use filtering technologies to mitigate infringement. One file-hosting site, 4shared, was purportedly unable to obtain fingerprint data from large rightsholder organizations for use in its implementation of the Echoprint system.

Finally, audio and video standards continue to evolve rapidly. For example, the H.264/MPEG-4 video coding standard has had 22 different versions since its inception in 2003; many of these version updates involve substantial updates to the coding function and format that enable more scalable video coding algorithms and faster processing rates. Similarly, the JPEG image compression format specification has had six different updates.
Audio and video streams can also be transcoded from one format to another, or from one bitrate to another. This process often distorts aspects of the video encoding that might be used for fingerprinting. For example, in the case of video compression, a common technique is to perform a process akin to static image compression on an “anchor frame,” with subsequent frames in the video encoded as differences from the anchor frame with motion vectors. This encoding process has slight differences between codecs; for example, the MPEG-4 standard defines forward motion vectors based on frames that are subsequent to an anchor frame; H.263, on the other hand, encodes motion vectors based on anchor frames that both precede and follow the frame being coded.

Transcoding between these formats fundamentally changes the nature of the encoding, given that the encoded motion vectors themselves will be different in each case. Transcoding between these formats fundamentally changes the nature of the encoding, given that the encoded motion vectors themselves will be different in each case. For example, because an MPEG-encoded video will have only forward motion vectors and an H.263 video will have motion vectors that refer both forwards and backwards in time in the video stream, converting between the two formats will fundamentally change the nature of how the underlying compressed video stream is represented. Similarly, audio codecs use different methods to encode the underlying waveform. One such method, called waveform encoding, tries to represent an analog waveform as accurately as possible with its digital equivalent; pulse code modulation (PCM) is an example of such an encoding. Other audio codecs use a method known as differential encoding (also known as predictive coding), which predicts a subsequent audio sample based on a previous one and stores only the difference between the predicted audio sample and the actual sample. Differential coding significantly reduces the bandwidth required for transmitting an audio signal; for example, G.729, a differential encoding, uses about half of the number of bits as G.711, a waveform encoding. Naturally, each of these encoding mechanisms produces very different digital representations of the same audio sample.

In general, the evolution and proliferation of standards increases the likelihood the underlying representation of the compressed video stream will not be preserved across formats. If the database containing a fingerprint is based on characteristics that are present in one format but absent in another, then the fingerprint will fail to

produce a match on the content. As standards and formats proliferate, so do the mechanisms for transcoding between them, each of which may manipulate the underlying format in ways that disrupt the attributes of the file that are used for fingerprinting.

As encoding and compression standards continue to evolve, the fingerprinting technologies themselves must also develop to keep pace with standards. In the case of an audio fingerprinting mechanism such as Echoprint or other audio fingerprinting technologies, the inter-onset interval features may be preserved across encoding formats because they are based on properties of the audio. On the other hand, if the technique is based on properties of the compressed encoded video stream, as discussed in the video compression example above, the fingerprinting mechanisms will fail as encoding mechanisms improve and evolve. At the same time, standards must continue to evolve to support the proliferation of digital content on the Internet; given that video is a significant fraction of all Internet traffic at peak, innovation in coding standards will improve the efficiency of compression as well as the resilience to errors, packet loss, and (ever increasing) congestion. Keeping standards fixed solely to facilitate fingerprinting will impede innovation and is not feasible or sensible.

IV. Negative Implications of Filtering Requirement

The technical limitations of filtering tools are reason enough to question the reasonableness of policy proposals that would require OSPs to implement such technologies. Given the limited types of content these tools can scan and the ease with which they can be circumvented, it is unrealistic to believe that use of filtering systems will have the type of impact on infringement that proponents would like. Of course, the merits of requiring OSPs to use content filtering tools must be evaluated in terms of their cost as well as their efficacy. The negative impact a proactive filtering mandate would have on OSPs—and the concomitant decrease in creative output from those who rely on the internet as a medium of production and distribution—significantly outweighs its benefits.
A. Filtering Tools Are Prohibitively Expensive for Many Small OSPs

Considering the cost of filtering technologies in the most literal sense—the economic cost to OSPs—reveals problems with injecting a filtering obligation into a law meant to promote investment. Contrary to claims made by proponents of mandatory filtering, such tools are quite expensive—particularly for the startups that have historically driven the growth of the internet sector. To support its proposed filtering mandate, the European Commission argued that the cost of filtering tools would be negligible for startups: “it is estimated that a small scale online service provider can obtain such services for less than 900 euros a month.”59 This estimate derives from comments Audible Magic submitted in a Copyright Office study of the effectiveness of the DMCA’s safe harbor provisions.60 But this estimate is only accurate for an incredibly small number of OSPs. Audible Magic’s website indicates that this price only applies to tools that filter audio files and is available only for OSPs hosting less than 5,000 song files per month—an incredibly low volume for an OSP.61 To put this in perspective, when Soundcloud was only five years old, users were uploading twelve hours of audio content every minute.62 The cost to most OSPs to source filtering technology from third parties is likely to be far more expensive in absolute terms than the EC’s projection (assuming such technologies even exist to scan the types of content such OSPs host). A recent survey of OSPs reported that medium-sized companies engaged in file-hosting services paid between $10,000 and $25,000 a month in licensing fees alone for Audible Magic’s filtering tool.63 It is worth noting that the licensing fees for the software amount to only a portion of the total costs associated with using fingerprinting software. Any OSP’s hosting platform must be altered or augmented to perform the fingerprint lookups and comparisons against a fingerprint database, a substantial software integration task.

59 Draft EC Report, Supra, note 41 at 138.
63 Urban, et al., supra note 53, at 64.
And, because not all OSPs that host “large amounts of content” (the undefined qualifying criteria for the EC’s filtering mandate) host large amounts of infringing content, requiring OSPs of a certain size to proactively filter user uploads will create massive costs for many OPSs without a commensurate benefit. For example, Kickstarter—a crowdfunding platform through which users can raise funds for creative projects—has hosted 338,368 projects on its site since its inception, most of which contain a range of image, audio, and video content; but in 2015, it removed only 78 projects in response to copyright infringement notices targeting 215 distinct projects. Because copyright infringements are exceedingly rare on its platform, the cost to Kickstarter of implementing filtering technologies that could process image, audio, and video content—likely separate tools for each media type—would far exceed the benefit of removing less than 100 infringements per year. And, since so many of the infringement notifications Kickstarter receives each year are false or deficient (around two-thirds of all infringement notices in 2015), it would likely still have to devote human resources to review the content that filtering technologies identified to prevent inappropriate removals.

This is not to say that filtering technologies are unaffordable for all OSPs. Indeed, many larger OSPs that allow users to share content already use filtering tools. YouTube famously spent $60 million developing its Content ID tool. Soundcloud spent more than €5 million building its own filtering technology and still must dedicate seven full-time employees to maintain the technology. But, for smaller OSPs, the cost of these systems would make it significantly harder to attract investors and compete with dominant incumbents. In a survey of investors in the U.S. and EU, a majority of respondents said they would be “uncomfortable investing in businesses that would be required by law to run a technological filter on user-uploaded content.” Since startups are responsible for virtually all new net job

---

64 Draft EC Report, Supra note 41, at 132.
67 Hearing on Section 512 of Title 17 before the H. Judiciary Subcomm. on Courts, Intellectual Prop., & the Internet, 113th Cong. 47 (2014), at 49 (testimony of Katherine Oyama).
68 https://ec.europa.eu/eusurvey/pdf/answer/6ac3b21-865a-402c-876a-e2b67c0ceef9.
69 Evan Engstrom et al., The Impact of Internet Regulation on Early Stage Investment, (2014), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/572a35e0b6a6d0fe01dec28/146238410881/EngineFifthEraCopyrightReport.pdf.
growth in the U.S., the overall economic cost of a mandatory filtering regime for OSPs would likely be significant.

B. A Filtering Obligation Would Undermine the Certainty of the Safe Harbor

Even if the monetary and labor expense of employing content filtering technology alone had a minimal direct impact on OSP viability, conditioning the safe harbor on implementing a satisfactory content filtering tool would likely curtail investment in OSPs by creating uncertainty as to whether an OSP’s chosen filtering technology would pass muster under the statute. A safe harbor is only useful for mitigating uncertainty if the prerequisites for obtaining its protections are clear, but it is decidedly unclear how policymakers could precisely define what sorts of filtering technologies would be deemed adequate to claim the safe harbor. Any filtering requirement would either have to endorse a particular technology (and quickly become outdated) or establish a “reasonableness” standard that would require clarification from courts before OSPs could have any confidence in their protection under the safe harbor.

Given the fast pace of development, attempting to identify in statute which particular filtering technologies would qualify for the safe harbor would be a fool’s errand. Since 2000, at least twelve new audio encoding formats have emerged, each of which has a plethora of encoding and sampling rates that can introduce distortion or loss. Given the rate of innovation in encoding and compression technologies, as well as the rate at which fingerprinting technologies themselves continue to evolve, mandating the use of a specific filtering technology is likely to be ineffective. Audio and video compression technologies typically are subject to significant updates every couple of years; requiring the use of a specific technology would most likely result in the use of technologies that quickly become obsolete and increasingly inaccurate as the standards and technologies themselves evolve.

71 Some commenters have proposed an even more stringent standard, arguing that OSPs should be required to implement the “best available” filtering technology to qualify for the safe harbor. See Helman & Parchomovsky, supra note 37.
Inevitably, a mandatory filtering regime would require a “reasonableness” standard for determining which filtering technologies are sufficient for an OSP to claim the safe harbor. Although “reasonableness” standards are common throughout the law—the DMCA itself conditions the safe harbor on an OSP “reasonably implement[ing]” a repeat infringer policy73—determining what filtering technologies might be considered “reasonable” is a particularly uncertain inquiry, considering the wide range of technologies that can be employed to filter content and the fast pace at which new technologies emerge and replace prior systems. While the “reasonableness” of an OSP’s repeat infringer policy is likely to remain consistent over time, a filtering technology that is reasonable for an OSP one year may be legally inadequate the next.

Even amongst existing filtering technologies, there is little agreement about which systems are sufficiently effective. For example, while some copyright owners have commended YouTube for developing its Content ID fingerprinting system, others argue that even Content ID—a technology that cost YouTube $60 million to develop—is inadequate and should be supplemented by some other, more effective system.74 Such complaints about the reasonableness and efficacy of filtering systems seem motivated in part by unrealistic expectations about what they can actually accomplish. According to Sony Music Entertainment, “if even a single copy of a particular work slips by a filter, in practical terms it is tantamount to a failure by the filter to lessen the availability of that work on the service at all.”75 Similarly, in the litigation over the infamous Napster peer-to-peer file sharing service, a judge held that Napster was in violation of an injunction because, despite filtering out 99.4 percent of infringing content, the court demanded the site use a filter that could remove all infringing material.76 Of course, since no such filtering technology exists, Napster promptly shut down.

Even determining what criteria courts should use in evaluating the reasonableness of a particular filtering system will likely produce significant uncertainty and

74 Compare Comments of the Motion Picture Association of America, supra note 35 (“voluntary cooperation between content owners and service providers could go a long way to solving this problem…[YouTube] uses a Content ID system that is very similar to those described in the UGC Principles”) with Comments of Universal Music Group, supra note 33 (“while YouTube’s content management tools were essential, and have improved to a certain extent today, they still are not sufficient.”).
75 Comments of Sony Music Group, U.S. Copyright Office, Section 512 Study, at 12-13 (2016).
potential conflicting judicial opinions: should the law favor systems that are capable of identifying a wide range of potentially infringing content at the risk of deleting a non-negligible amount of non-infringing content (such as metadata searches) or should it prioritize accuracy at the risk of underinclusiveness, as with hash-based filtering? Is a system that is technically state-of-the-art “reasonable” for purposes of a filtering mandate if copyright owners refuse to submit fingerprints for that system?77

Ultimately, since what is “reasonable” is generally a question of fact for a jury to decide,78 an OSP would likely not know the answer to these questions—and consequently whether it could claim the safe harbor—until after a full jury trial, rendering the safe harbor of little value. For many OSFs, simply engaging in litigation over the adequacy of a filtering system would lead to bankruptcy, even with a favorable judgment on the merits.79 The threat of protracted litigation would likely deter investment in startups seeking to compete with incumbent content distribution OSFs, since venture investors would have to consider whether a potential portfolio company could afford to implement a content filtering technology and defend a secondary infringement lawsuit. The likelihood of decreased venture activity is not merely speculative, as the negative investment consequences associated with increased litigation risk are well-documented.80 In short, if the DMCA’s safe harbor protections were predicated on an OSP implementing a “reasonable” content filtering system, the uncertainty surrounding any OSP’s eligibility for the safe harbor would undermine one of the central purposes of the DMCA: providing “more certainty ... in order to [allow OSFs to] attract the substantial investments necessary to continue the expansion and upgrading of the Internet.”81

---

77 See, e.g., Ernesto, 4shared’s Piracy “Fingerprint” Tool Helps to Reduce Takedown Notices, TORRENTFREAK (Nov. 23, 2016), https://torrentfreak.com/4shared-copyright-holders-abuse-googles-dmca-takedown-system-161123/ (“While 4shared has been using the content recognition software for quite a while already, not all copyright holders are eager to use it. Several large industry groups such as IFPI refuse to provide 4shared with fingerprint data.”).
78 See Perfect 10, Inc. v. CCBill LLC, 488 F.3d 1102 (9th Cir. 2007) (reasonableness of repeat infringer policy a question of fact); see also Tran v. State Farm Mut. Auto. Ins. Co., 999 F. Supp. 1369, 1372 (D. Haw. 1998) (“An analysis of what is reasonable is almost always de facto a question for the jury.”).
79 For example, Veoh, a video hosting OSP, was forced to declare bankruptcy after its protracted and ultimately successful defense of a secondary infringement lawsuit brought by Universal Music Group. Eliot van Buskirk, Veoh Files For Bankruptcy After Fending Off Infringement Charges, WIRED (February 12, 2010) https://www.wired.com/2010/02/veoh-files-for-bankruptcy-after-fending-off-infringement-charges/.
81 Hatch, supra note 6.
C. A Filtering Obligation Would Likely Have Minimal Impact on Infringement

Though the DMCA does not require OSPs to employ content filtering technologies, many of the larger OSPs that can afford to do so voluntarily use some kind of content filtering,82 often because they are concerned that failing to do so could result in secondary liability (though as described in more detail supra, some OSPs have implemented filtering tools pursuant to cooperative agreements to help copyright owners monetize user posts).83 As such, passing a law requiring OSPs to filter content would not result in any decreased infringing activity on many of the larger OSPs that already use filtering tools, such as YouTube, Soundcloud, and Spotify. And, even assuming that there are large numbers of OSPs that would implement filtering tools for the first time should such a requirement become law, and assuming that these tools were effective at identifying and removing infringing content, requiring U.S.-based OSPs to implement such tools to obtain the safe harbor would likely not have a major impact on copyright infringement, since most infringing activity appears to involve so-called “pirate sites” that are based overseas or are otherwise ignoring U.S. copyright laws. One recent study analyzing the interplay between online advertising and copyright infringement identified the top 589 “pirate” sites from 2014, virtually all of which are foreign domains.84 Similarly, all of the sites identified in a report from TorrentFreak on the ten most popular torrent85 sites of 2017 appear to be located outside the U.S.86 While it is impossible to know with certainty where some websites are located, the Motion Picture Association of America’s 2016 public commentary submission in an inquiry from the Office of the U.S. Trade Representative regarding “notorious markets” purports to identify the locations of many such sites, all of which are based outside

---

82 There is some debate about the prevalence of content filtering technologies amongst OSPs. One survey of OSPs concluded that “most OSPs do not filter,” while other commentators have suggested that certain forms of filtering are common, at least amongst larger OSPs. Compare Urban, et al., supra note 53 at 58 with Gallo, supra note 26 at 296 (“[F]ingerprinting technology is widespread among UGCs”).

83 See Urban, supra note 53 at 58 (“OSPs reported that filtering is likely to be adopted under considerable pressure and concern for liability.”).

84 Digital Citizens Alliance, Good Money Still Going Bad: Digital Thieves and the Hijacking of the Online Ad Business (2015) http://illusionofmore.com/wp-content/uploads/2015/05/latest-DigitalCitizensAlliance5.pdf. Moreover, of the top 596 "pirate" sites identified in the 2013 report, less than 60 percent made the list in 2014, indicating that these sites, regardless of domicile, are too transient to be impacted by anti-copyright infringement measures.

85 Torrent technology is content-neutral and can be used to distribute non-infringing and infringing content alike.

Requiring OSPs to filter user uploads would do nothing to stop infringement on these websites.

V. Conclusion

In the two decades since the DMCA was passed, internet activity and total creative output have both increased exponentially. Considering the DMCA was designed to mitigate legal uncertainty in order to drive investment to the internet sector and to encourage copyright owners and OSPs to cooperate to combat online infringement in order to promote creative production, its success seems unquestionable. But, despite this strong record, the DMCA has come under fire in recent years from copyright industries for allegedly not doing enough to stop copyright infringement. These calls to modify the DMCA to require OSPs to implement technical measures to police infringement overestimate the technical capacity of these tools and underestimate the damage that their obligatory use would have on the internet ecosystem. Before considering dangerous mandatory content filtering rules, policymakers should understand the inherent limitations of filtering technologies. Reversing two decades of sensible copyright policy to require OSPs to deploy tools that are costly, easily circumvented, and limited in scope would deeply harm startups, users, and content creators alike.

Engine (www.engine.is) is a non-profit organization that supports the growth of technology entrepreneurship through economic research, policy analysis, and advocacy on local and national issues.
This is what Google says search will look like under EU copyright laws

Look what EU made me do

By James Vincent  Jan 17, 2019, 12:26pm EST

Last September, the European Parliament voted in favor of the Copyright Directive: a sweeping piece of legislation intended to update copyright for the internet age, but critics said it would fundamentally break the internet. Now, as negotiations over the directive's final wording draw to a close, Google has issued a warning about the damage the directive might do in an unusual format: an empty search results page.

To be more precise, it’s an empty search results page for news. One of the most controversial segments of the Copyright Directive is Article 11, which gives publishers the right to demand paid licenses for using snippets of their stories. From Google's point of view, that gives it two choices: start paying for licenses or don’t show snippets at all.

In “test” screenshots (first shared with Search Engine Land) the tech giant demonstrated what the latter approach might look like. If a user in the EU searches for “latest news,” they would simply see links to media outlets’ sites alongside some tantalizingly useless timestamps. No summaries of stories, no headlines or pictures — no nothing.
In a blog post published last December, Google’s VP of News, Richard Gingras, explains the company’s position on Article 11, arguing that the benefits it might deliver (e.g., license payments to newspapers) would be skewed toward larger publishers, while Google would be forced to pick “winners and losers” when paying for content.

“Online services, some of which generate no revenue (for instance, Google News) would have to make choices about which publishers they’d do deals with,” Gingras wrote. “Presently, more than 80,000 news publishers around the world can show up in Google News, but Article 11 would sharply reduce that number.”

Diego Naranjo, senior policy advisor at digital rights group EDRi, told The Verge that Google was likely releasing these screenshots as a scare tactic, but that they weren’t an unreasonable interpretation of the legislation. “They’re just trying to show what Article 11 will push them to do,” said Naranjo. “And in that sense, this is a possibility.”

**READ MORE:** [WHAT YOU NEED TO KNOW ABOUT THE EU COPYRIGHT DIRECTIVE](https://www.theverge.com/2019/1/1/119/google-eu-copyright-laws-search-news-link-tax)

And what about the Copyright Directive itself? How close is that to being finalized? As is often the case with EU legislation, there is no clear answer. After the European Parliament voted in favor of the directive last September, it entered a process known as “trilogues,” which are closed-door negotiations between the European Commission, Council, and Parliament.

These trilogues give EU politicians (MEPs) a chance to tweak the wording of the Copyright Directive and potentially remove some troublesome parts. To do so, representatives from countries making up more than 36 percent of the EU’s population would have to band together. This is why bodies like the [Electronic Frontier Foundation](https://www.eff.org) have been encouraging citizens in key countries like Poland and Germany to lobby their representatives.

But these negotiations are not public, and it’s impossible to predict what changes, if any, might be made. The next major milestone will be a vote on January 18th by the European Council. If the legislation passes, this there will be a final negotiation between the European Parliament and Council on January 21st, and then a final vote sometime in March. After that, EU nations will have to start implementing the Copyright Directive into national law in 2021, and Google’s bare search results could become a reality.
Accountable Algorithms

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu†

Abstract

Many important decisions historically made by people are now made by computers. Algorithms count votes, approve loan and credit card applications, target citizens or neighborhoods for police scrutiny, select taxpayers for IRS audit, grant or deny immigration visas, and more.

The accountability mechanisms and legal standards that govern such decision processes have not kept pace with technology. The tools currently available to policymakers, legislators, and courts were developed to oversee human decision makers and often fail when applied to computers instead. For example, how do you judge the intent of a piece of software? Because automated decision systems can return potentially incorrect, unjustified or unfair results, additional approaches are needed to make such systems accountable and governable. This Article reveals a new technological toolkit to verify that automated decisions comply with key standards of legal fairness.

We challenge the dominant position in the legal literature that transparency will solve these problems. Disclosure of source code is often neither necessary (because of alternative techniques from computer science) nor sufficient (because of the issues analyzing code) to demonstrate the fairness of a process. Furthermore, transparency may be undesirable, such as when it permits tax cheats or terrorists to game the systems determining audits or security screening or discloses private information.

The central issue is how to assure the interests of citizens, and society as a whole, in making these processes more accountable. This Article argues that technology is creating new opportunities—subtler and more flexible than total transparency—to design decisionmaking algorithms so that they better align with legal and policy objectives. Doing so will improve not only the current governance of automated decisions, but also—in certain cases—the governance of decisionmaking in general. The implicit (or explicit) biases of human decisionmakers can be difficult to find and root out, but we can peer into the “brain” of an algorithm: computational processes and purpose specifications can be declared prior to use and verified afterwards.

The technological tools introduced in this Article apply widely. They can be used in designing decisionmaking processes from both the private and public sectors, and they can be tailored to verify different characteristics as desired by decisionmakers, regulators, or the public. By forcing a more careful consideration of the effects of decision rules, they also engender policy discussions and closer looks at legal standards. As such, these tools have far-reaching implications throughout law and society.

† Respectively, System Engineer, CloudFlare and Affiliate, Center for Information Technology Policy, Princeton; Associate Director, Center for Information Technology Policy, Princeton; Post Doctoral Research Associate, Princeton; Robert E. Kahn Professor of Computer Science and Public Affairs, Princeton; Stanley D. and Nikki Waxberg Chair in Law, Fordham Law School; Principal, Upturn; Principal, Upturn. For helpful comments, the authors are very grateful to participants at the Berkeley Privacy Law Scholars Conference and at the NYU School of Law conference on “Accountability and Algorithms.”
Part I of this Article provides an accessible and concise introduction to foundational computer science concepts that can be used to verify and demonstrate compliance with key standards of legal fairness for automated decisions without revealing key attributes of the decisions or the processes by which the decisions were reached. Part II then describes how these techniques can assure that decisions are made with the key governance attribute of procedural regularity, meaning that decisions are made under an announced set of rules consistently applied in each case. We demonstrate how this approach could be used to redesign and resolve issues with the State Department’s diversity visa lottery. In Part III, we go further and explore how other computational techniques can assure that automated decisions preserve fidelity to substantive legal and policy choices. We show how these tools may be used to assure that certain kinds of unjust discrimination are avoided and that automated decision processes behave in ways that comport with the social or legal standards that govern the decision. We also show how automated decision making may even complicate existing doctrines of disparate treatment and disparate impact, and we discuss some recent computer science work on detecting and removing discrimination in algorithms, especially in the context of big data and machine learning. And lastly, in Part IV, we propose an agenda to further synergistic collaboration between computer science, law, and policy to advance the design of automated decision processes for accountability.

Table of Contents
I. How Computer Scientists Build and Evaluate Software .................................................. 10
   A. Assessing Computer Systems ...................................................................................... 10
      1. Static Analysis: Review from Source Code Alone ................................................... 13
      2. Dynamic Testing: Examining a Program’s Actual Behavior ...................................... 16
      3. The Fundamental Limit of Testing: Noncomputability ............................................. 18
   B. The Importance of Randomness ................................................................................... 19
II. Designing Computer Systems for Procedural Regularity ............................................... 22
   A. Transparency and Its Limits ...................................................................................... 23
   C. Technical Tools for Procedural Regularity ................................................................ 27
      1. Software Verification .............................................................................................. 27
      2. Cryptographic Commitments ................................................................................. 30
      3. Zero-Knowledge Proofs ......................................................................................... 32
      4. Fair Random Choices ............................................................................................. 33
   D. Applying Technical Tools Generally .......................................................................... 36
      1. Current DVL Procedure .......................................................................................... 38
      2. Transparency Is Not Enough .................................................................................. 39
      3. Designing the DVL for Accountability ..................................................................... 39
III. Designing Algorithms to Assure Fidelity to Substantive Policy Choices ......................... 41
A. Machine Learning, Policy Choices, and Discriminatory Effects ................................................. 42
B. Technical Tools for NonDiscrimination ......................................................................................... 45
   1. Learning from Experience ........................................................................................................... 46
   2. Fair Machine Learning ............................................................................................................. 47
   3. Discrimination, Data Use, and Privacy .................................................................................... 51
C. AntiDiscrimination Law and Algorithmic Decisionmaking .......................................................... 53
   2. Ricci Impels Designing for NonDiscrimination ...................................................................... 56
IV. Fostering Collaboration Across Computer Science, Law, and Policy ........................................... 56
A. Recommendations for Computer Scientists: Design for After-the-Fact Oversight .................. 57
B. Recommendations for Lawmakers and Policymakers .................................................................. 60
   1. Reduced Benefits of Ambiguity ................................................................................................. 60
   2. Accountability to the Public ....................................................................................................... 63
   3. Secrets and Accountability ........................................................................................................ 65

Introduction

Many important decisions that were historically made by people are now made by computer systems: ¹ votes are counted; voter rolls are purged; loan and credit card applications are approved; ² welfare and financial aid decisions are made; ³ taxpayers are chosen for audits; citizens or neighborhoods are targeted for

---

¹ In this work, we use the term “computer system” where others have used the term “algorithm.” See, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015). This allows us to separate the concept of a computerized decision from the actual machine that effects it. See infra note 14 for a more detailed explanation.
police scrutiny;\(^4\) air travelers are selected for search;\(^5\) and visas are granted or denied. The efficiency and accuracy of automated decisionmaking ensures that its domain will continue to expand. Even mundane activities now involve complex computerized decisions: everything from cars to home appliances now regularly executes computer code as part of its normal operation.

However, the accountability mechanisms and legal standards that govern decision processes have not kept pace with technology. The tools currently available to policymakers, legislators, and courts were developed primarily to oversee human decisionmakers. Many observers have argued that our current frameworks are not well adapted for situations in which a potentially incorrect,\(^6\) unjustified,\(^7\) or unfair\(^8\) outcome emerges from a computer. Citizens, and society as a whole, have an interest in making these processes more accountable. If these new inventions are to be made governable, this gap must be bridged.

In this Article, we describe how authorities can demonstrate--and how the public at large and oversight bodies can verify--that automated decisions comply with key standards of legal fairness. We consider two approaches: ex ante approaches aiming to establish that the decision process works as expected (which are commonly studied by technologists and computer scientists), and ex post approaches once decisions have been made, such as review and oversight (which are common in existing governance structures). Our proposals aim to use the tools of the first approach to guarantee that the second approach can function effectively. Specifically, we describe how technical tools for verifying the correctness of computer systems can be used to ensure that appropriate evidence exists for later oversight.

We begin with an accessible and concise introduction to the computer science concepts on which our argument relies, drawn from the fields of software verification, testing, and cryptography. Our argument builds on the fact that technologists can and do verify, for themselves, that software systems work in accordance with designs. No computer system is built and deployed in the

---


\(^5\) See Notice of Modified Privacy Act System of Records, 78 Fed. Reg. 55,270, 55,271 (Sept. 10, 2013) (“[T]he passenger prescreening computer system will conduct risk-based analysis of passenger data . . . . TSA will then review this information using intelligence-driven, risk-based analysis to determine whether individual passengers will receive expedited, standard, or enhanced screening . . . .”).


\(^7\) See id., at 1256-57 (noting the “crudeness” of algorithms designed to identify potential terrorists that yield a high rate of false positives).

\(^8\) See Solon Barocas & Andrew D. Selbst, Big Data’s Disparate Impact, 104 Calif. L. Rev. 671, 677 (2016) (“[D]ata mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage.”).
world shrouded in total mystery. While we do not advocate any specific liability regime for the creators of computer systems, we outline the range of tools that computer scientists and other technologists already use, and show how those tools can ensure that a system meets specific policy goals. In particular, while some of these tools provide assurances only to the system’s designer or operator, other established methods could be leveraged to convince a broader audience, including regulators or even the general public.

The tools available during the design and construction of a computer system are far more powerful and expressive than those that can be bolted on to an existing system after one has been built. We argue that in many instances, designing a system for accountability can enable stakeholders to reach accountability goals that could not be achieved by imposing new transparency requirements on existing system designs.

We show that computer systems can be designed to prove to oversight authorities and the public that decisions were made under an announced set of rules consistently applied in each case, a condition we call procedural regularity. The techniques we describe to ensure procedural regularity can be extended to demonstrate adherence to certain kinds of substantive policy choices, such as blindness to a particular attribute (e.g., race in credit underwriting). Procedural regularity ensures that a decision was made using consistently applied standards and practices. It does not, however, guarantee that such practices are themselves good policy. Ensuring that a decision procedure is well justified or relies on sound reasoning is a separate challenge from achieving procedural regularity. While procedural regularity is a well understood and generally desirable property for automated and nonautomated governance systems alike, it is merely one principle around which we can investigate a system’s fairness.

It is common, for example, to ask whether a computer system avoids certain kinds of unjust discrimination, even when such systems are blind to certain attributes (e.g., gender in automated hiring decisions). We later expand our discussion and show how emerging computational techniques can assure that automated decisions satisfy other notions of fairness that are not merely procedural, but actively consider a system’s effects. We describe in particular detail techniques for avoiding discrimination, even in machine learning systems that derive their decision rules from data rather than from code written by a programmer. Finally, we propose next steps to further the emerging and critically important collaboration between computer scientists and policymakers.

---

9 Although some machine-learning systems produce results that are difficult to predict in advance and well beyond traditional interpretation, the choice to field such a system instead of one which can be interpreted and governed is itself a decision about the system’s design. While we do not advocate that any approach should be forbidden for any specific problem, we aim to show that advanced tools exist that provide the desired functionality while also permitting oversight and review.
Legal scholars have argued for twenty years that automated processing requires more transparency,¹⁰ but it is far from obvious what form such transparency should take. Perhaps the most obvious approach is to disclose a system’s source code, but this is at best a partial solution to the problem of accountability for automated decisions. The source code of computer systems is illegible to nonexperts. In fact, even experts often struggle to understand what software code will do: inspecting source code is a very limited way of predicting how a computer program will behave.¹¹ Machine learning, one increasingly popular approach to automated decisionmaking, is particularly ill-suited to source code analysis because it involves situations where the decisional rule itself emerges automatically from the specific data under analysis, sometimes in ways that no human can explain.¹² In this case, source code alone teaches a reviewer very little, since the code only exposes the machine learning method used and not the data-driven decision rule.

Moreover, in many of the instances that people care about, full transparency will not be possible. The process for deciding which tax returns to audit, or whom to pull aside for secondary security screening at the airport, may need to be partly opaque to prevent tax cheats or terrorists from gaming the system. When the decision being regulated is a commercial one, such as an offer of credit, transparency may be undesirable because it defeats the legitimate protection of consumer data, commercial proprietary information, or trade secrets. Finally, when an explanation of how a rule operates requires disclosing the data under analysis and those data are private or sensitive (e.g., in adjudicating a commercial offer of credit, a lender reviews detailed financial information about the applicant), disclosure of the data may be undesirable or even legally barred.

Furthermore, making the rule transparent—whether through source code disclosure or otherwise—may still fail to resolve the concerns of many participants. No matter how much transparency surrounds a rule, people can still wonder whether the disclosed rule was actually used to reach a decision in their own cases. Particularly where an element of randomness is involved in the process, a person audited or patted down may wonder: Was I really chosen by the rule, or has some bureaucrat singled me out on a whim? But full disclosure of how particular decisions were reached is often unattractive because the decisions themselves often incorporate sensitive health, financial, or other private information either as input,

---

¹⁰ See, e.g., Citron, supra note 6, at 1253 (describing automated decisionmaking as “adjudicat[ion] in secret”); Paul Schwartz, Data Processing and Government Administration: The Failure of the American Legal Response to the Computer, 43 Hastings L.J. 1321, 1323-25 (1992) (“So long as government bureaucracy relies on the technical treatment of personal information, the law must pay attention to the structure of data processing . . . . There are three essential elements to this response: structuring transparent data processing systems; granting limited procedural and substantive rights . . . and creating independent governmental monitoring of data processing systems.” (emphasis omitted)).
¹¹ See infra subsection I.A.1 (discussing static analysis).
output, or both (for example, an individual’s tax audit status may be sensitive or protected on its own, but it may also imply details about that individual’s financial data).

Even full disclosure of a decision’s provenance to that decision’s subject can be problematic. Most individuals are ill-equipped to review how computerized decisions were made, even if those decisions are reached transparently. Further, the purpose of computer-mediated decisionmaking is to bring decisions an element of scale, where the same rules are ostensibly applied to a large number of individual cases or are applied extremely quickly. Individuals auditing their own decisions (or experts assisting them) would be both inundated with the need to review the rules applied to them and often able to generalize their conclusions to the results of others, raising the same disclosure concerns described above. That is, while transparency of a rule makes reviewing the basis of decisions more possible, it is not a substitute for individualized review of particular decisions.\(^\text{13}\)

Fortunately, technology is creating new opportunities--more subtle and flexible than total transparency--to make automated decisionmaking more accountable to legal and policy objectives. Although the current governance of automated decisionmaking is underdeveloped, computerized processes can be designed for governance and accountability. Doing so will improve not only the current governance of computer systems, but also--in certain cases--the governance of decisionmaking in general.

This Article argues that in order for a computer system to function in an accountable way--either while operating an important civic process or merely engaging in routine commerce--accountability must be part of the system’s design from the start. Designers of such systems--and the nontechnical stakeholders who often oversee or control system design--must begin with oversight and accountability in mind. We offer examples of currently available tools that could aid in that design, as well as suggestions for dealing with the apparent mismatch between policy ambiguity and technical precision.

In Part I of this Article, we provide an accessible introduction to how computer scientists build and evaluate computer systems and the software and algorithms\(^\text{14}\) that comprise them. In particular, we describe how computer scientists

\(^{13}\) Even when experts can pool investigative effort across many decisions, there is no guarantee that the basis for decisions will be interpretable or that problems of fairness or even overt special treatment for certain people will be discovered. Further, a regime based on individuals auditing their own decisions cannot adequately address departures from an established rule, which favor the individual auditing her own outcome, or properties of the rule, which can only be examined across individuals (such as nondiscrimination).

\(^{14}\) In this Article, we limit our use of the word “algorithm” to its usage in computer science, where it refers to a well-defined set of steps for accomplishing a certain goal. In other contexts, where other authors have used the term “algorithm,” we describe “automated decision processes” reflecting “decision policies” implemented by pieces of “software,” all comprising “computer systems.” Our adoption of the phrase “computer systems” was suggested by (and originally due to) Helen Nissenbaum and we are grateful for the precision it provides. See generally Batya Friedman & Helen Nissenbaum, \textit{Bias in Computer Systems}, 14 ACM Transactions on Info. Sys. 330 (1996).
evaluate a program to verify that it has desired properties, and discuss the value of randomness in the construction of many computer systems. We characterize what sorts of properties a computer system can be tested and describe one of the fundamental truths of computer science—that there are some properties of computer systems which cannot be tested completely. We observe that computer systems fielded in the real world are (or at least should be) tested regularly during creation, deployment, and operation, merely to establish that they are actually functional.

Part II examines how to design computer systems for procedural regularity, a key governance principle enshrined in law and public policy in many societies. We consider how participants, decision subjects, and observers can be assured that each individual decision was made according to the same procedure—for example, how observers can be assured that the decisionmaker is not choosing outcomes on a whim while merely claiming to follow an announced rule. We describe why mere disclosure of a piece of source code can be impractical or insufficient for these ends. Indeed, without full transparency—including source code, input data, and the full operating environment of the software—even the disclosure of audit logs showing what a program did while it was running provides no guarantee that the disclosed

The term “algorithm” is assigned disparate technical meanings in the literatures of computer science and other fields. The computer scientist Donald Knuth famously defined algorithms as separate from mathematical formulae in that (1) they must “always terminate after a finite number of steps;” (2) “[e]ach step of an algorithm must be precisely defined; the actions to be carried out must be rigorously and unambiguously specified for each case;” (3) input to the algorithm is “quantities that are given to it initially before the algorithm begins;” (4) an algorithm’s output is “quantities that have a specified relation to the inputs;” and (5) the operations to be performed in the algorithm “must all be sufficiently basic that they can in principle be done exactly and in a finite length of time by someone using pencil and paper.” 1 Donald E. Knuth, The Art of Computer Programming: Fundamental Algorithms 4-6 (1968). Similarly and more simply, a widely used computer science textbook defines an algorithm as “any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output.” Thomas H. Cormen et al., Introduction to Algorithms 10 (2d ed. 2001).

By contrast, communications scholar Christian Sandvig says that “‘algorithm’ refers to the overall process” by which some human actor uses a computer to do something, including decisions made by humans as to what the computer should do, choices made during implementation, and even choices about how algorithms are represented and marketed to the public. Christian Sandvig, Seeing the Sort: The Aesthetic and Industrial Defense of “The Algorithm,” Media-N, http://median.newmediacaucus.org/art-infrastructures-information/seeing-the-sort-the-aesthetic-and-industrial-defense-of-the-algorithm [https://perma.cc/29E4-S44S]. Sandvig argues that even algorithms as simple as sorting “have their own public relations” and are inherently human in their decisions. Id.

Another communications scholar, Nicholas Diakopoulos, defines algorithms in the narrow sense as “a series of steps undertaken in order to solve a particular problem or accomplish a defined outcome,” but also considers them in the broad sense, saying that “algorithms can arguably make mistakes and operate with biases,” which does not make sense for the narrower technical definition. Nicholas Diakopoulos, Algorithmic Accountability: Journalistic Investigation of Computational Power Structures, 3 Digital Journalism 398, 398, 400 (2015). This confusion is common to much of the literature on algorithms and accountability, which we describe throughout this paper.

To avoid confusion, this paper adopts the precise definition of the word “algorithm” from computer science and, following Friedman and Nissenbaum, refers to the broader concept of an automated system deployed in a social or human context as a “computer system.”
information actually reflects a computer system’s behavior.\textsuperscript{15} In order to move beyond the need for full transparency, we focus on tools that can communicate partial information about secret processes, so that accountability and oversight continue to function even when policy interests, personal privacy, trade secrets, or other concerns protect a computer system, a piece of software, its inputs, its outputs, or its environment from disclosure. Putting it all together, we provide an illustrative example of how to redesign an existing, legally mandated automated decisionmaking system--the State Department’s Diversity Visa Lottery--so that it is provably accountable.

Part III considers the broader question of how to assess a computer system’s compliance with policy principles that go beyond procedural regularity. These broader properties include determining whether automated decision systems treat people (including protected groups) in ways that comport with the social or legal standards that govern the decision being made.\textsuperscript{16} This broadening raises the issue of translating a policy principle into a property of the system. Certain substantive policy choices translate easily: for example, prespecified rules such as blindness to a sensitive attribute.\textsuperscript{17} However, defining other policy objectives, such as a general notion of nondiscrimination, is a more complicated and fraught affair, particularly when systems rely on machine learning rather than decision rules explicitly predetermined by humans. We explore in particular the discriminatory effect that automated decisionmaking can have, noting real-world examples of newfound risks and describing some system properties that may align with policy goals. Finally, we observe how automated decisionmaking may complicate the existing doctrine of disparate treatment and disparate impact.

Part IV concludes by calling for increased collaboration between computer scientists and policymakers to develop and apply technical tools for the governance of computer systems. Given the ever-widening reach of automated decisions, computer scientists need to understand the policy challenges of oversight, and policymakers need to understand where new and emerging software tools can help address those challenges. We offer recommendations for bridging the gap between technologists’ desire for specificity and the policy process’s need for ambiguity. As a first step, we urge policymakers to recognize that accountability is feasible even when the details of a computer system are not fully known or must be kept secret.

\textsuperscript{15} The environment of a computer system includes anything it might interact with. For example, an outside observer will need to know what other software was running on a particular computer to ensure that nothing modified the behavior of the disclosed program. Some programs also observe (and change their behavior based on) the state of the computer they are running on (such as which files were or were not present or what other programs were running), the time they were run, or even the configuration of hardware on the system on which they were run.

\textsuperscript{16} This type of evaluation depends upon having already verified procedural regularity: if it cannot be determined that a particular algorithm was used to make a decision, it is fruitless to try to verify properties of that algorithm.

\textsuperscript{17} A concrete example would be the requirement that a decision only account for certain information for certain purposes, as in a system for screening job applicants that is allowed to take the gender of applicants as input, but only for the purpose of keeping informational statistics and not for making screening decisions.
We also argue that the ambiguities, contradictions, and uncertainties of the policy process need not discourage computer scientists from engaging constructively in it.

I. How Computer Scientists Build and Evaluate Software

Fundamentally, computers are general purpose machines that can be programmed to do any computational task, though they lack the desirable specificity and limitations of physical devices. Engineers often seek strong digital evidence that a computer system is working as intended. Such evidence may be persuasive for the system’s creator or operator only, for a predesignated group of receivers such as an oversight authority, or for the public at large. In many cases, systems are carefully evaluated and tested before they make it to the real world. Evidence that is convincing to the public and sufficiently nonsensitive to be disclosed widely is the most effective and desirable for ensuring accountability.

In this Part, we examine how computer scientists think about software assurance, how software is built and tested in the software industry, and what tools are available to get assurances about an individual piece of software or a large computer system. Thus, this section provides a brief and accessible map of key concepts and offers some insight into how computer scientists think about and approach these challenges.

A. Assessing Computer Systems

In general, a computer program is something that takes a set of inputs and produces a set of outputs. All too often, programs fail to work as their authors intended, because the programs have bugs or make assumptions about the input data which are not always true. Programmers often structure or design programs with an eye towards evaluation and testing in order to avoid or minimize these pitfalls. Many respected and popular approaches to software engineering are based on the idea that code should be written in ways that make it easier to analyze. For example, the programmer can:

- Organize the code into modules that can be evaluated separately and then combined.
- Test these modules for proper functionality both individually and in groups, possibly even testing the entire computer system end-to-end. Such testing generally involves writing test cases, or expected scenarios in which each module

---

18 For example, hydraulically operated control surfaces in a vehicle will telegraph resistance to the operator when they are close to a dangerous configuration, but the same controls operated by computer can omit feedback, allowing the computer to request configurations of actuators that are beyond their tolerances. This is a problem especially in the design of robotic arms and fly-by-wire systems for aircraft.


20 In particular, Test Driven Development (TDD) is a software engineering methodology practiced by many major software companies. For a general description of how TDD integrates automated testing into software design, see Kent Beck, Test-Driven Development: By Example (2003).

21 See Hunt & Thomas, supra note 19, at 1.
will run, and may involve running test cases each time the software is changed to avoid introducing new bugs or taking away functionality unintentionally.\footnote{See Steve McConnell, Code Complete: A Practical Handbook of Software Construction (2d ed. 2004).}

- Annotate the code with “assertions,” simple statements about the code that describe error conditions under which the program should crash immediately. Assertions are intended to be true if the program is running as expected. They become false when something is amiss and cause the program to crash (with an error message), rather than continuing in an errant state. In this way, assertions are a special kind of program error—a point at which a piece of software considers its internal state and its environment, and stops if these do not match what had been assumed by the program’s author.\footnote{This technique is originally due to I Herman H. Goldstine & John von Neumann, Planning and Coding of Problems for an Electronic Computing Instrument (1947), but it is now a widely used technique. For a historical perspective, see Lori A. Clarke & DavidS. Rosenblum, A Historical Perspective on Runtime Assertion Checking in Software Development, Software Engineering Notes (ACM Special Interest Grp. on Software Eng’g, New York, N.Y.), May 2006, at 25.}

- Provide a detailed description specifying the program’s behavior along with a machine-checkable proof that the code satisfies this specification. This differs from using assertions in that the proof guarantees ahead of time that the program will work as intended in all cases (or, equivalently, that an assertion of the facts covered by the proof will never fail to be true). When feasible, this approach is the most helpful thing a programmer can do to facilitate testing because it can provide real proof (rather than just circumstantial evidence or evidence linked to a particular point in a program’s execution, as with assertions) that the whole program operates as expected.\footnote{A simple example is a technique called model checking, usually applied to computer hardware designs, in which the property desired and the hardware or program are represented as logical formulae, and an automated tool performs an exhaustive search (i.e., tries all possible inputs) to check whether those formulae are not consistent. See, e.g., Edmund M. Clarke, Jr. et al., Model Checking (1999). An even simpler example comes from the concept of \textit{types} in programming languages, which associate the data values on which the program operates into descriptive classes and provide rules for how those classes should interact. For example, it should not be possible to add mathematically a number like “42” to a string of text like “Hello, World!” Because both kinds of data are represented inside the computer as bits and bytes, without a type system, the computer would be free to try executing this nonsensical behavior, which might lead to bugs. Type systems can help programmers avoid mistakes and express extremely complex relationships among the data processed by the program. For a more thorough explanation of type systems and model checking, see Benjamin C. Pierce, Types and Programming Languages (2002).}

These techniques are illustrative examples from a larger toolbox. Testing and verification of software and the development of tools to facilitate it comprise a rich and active subfield of computer science research. A thriving industry builds tools to assist in the development of software, and there is a constant debate about the practice of software engineering in the technology industry more broadly. Below, we give a brief taxonomy of this area, as well as some examples of tools, expectable results, and limitations of each category. We conclude this Section by
describing why testing code after it has been written, however extensively, cannot provide true assurance of how the system works, because any analysis of an existing computer program is inherently and fundamentally incomplete. This incompleteness implies that observers can never be certain that a computer system has a desired property unless that system has been designed to guarantee that property.

When technologists evaluate computer systems, they attempt to establish invariants, or facts about a program’s behavior which are always true regardless of a program’s internal state or what input data the program receives.25 Invariants can cover details as small as the behavior of a single line of code but can also express complex properties of entire programs, such as which users have access to which data or under what conditions the program can crash. By structuring code modules and programs in a way that makes it easy to establish simple invariants, it is possible to build an entire computer system for which important desirable invariants can be proved.26 Together, the set of invariants that a program should have make up its specification.27

Software code is, ultimately, a rigid and exact description of itself: the code both describes and causes the computer’s behavior when it runs.28 In contrast, public policies and laws are characteristically imprecise, often deliberately so.29 Thus, even when a well-designed piece of software does assure certain properties, there will always remain some room to debate whether those assurances match the

26 See id. at 1.
27 Specifications can be formal and written in a specification language, in which case they are rather like computer programs unto themselves. For example, the early models of core internet technology were written in a language called LOTOS, built for that purpose. See Tommaso Bolognesi & Ed Brinksma, Introduction to the ISO Specification Language LOTOS, 14 Computer Networks & ISDN Sys. 25, 1 (1987). Other common specification languages in practical use include Z and UML. It is even possible to build an executable program by compiling such a specification into a programming language or directly into machine code, an area of computer science research known as program synthesis. See Zohar Manna & Richard Waldinger, A Deductive Approach to Program Synthesis, 2 ACM Transactions on Programming Languages & Sys. 90, 1 (1980). Research has shown that when the language in which a program or specification is written more closely matches a human-readable description of the program’s design goals, programs are written with fewer bugs. See Michael C. McFarland et al., The High-Level Synthesis of Digital Systems, 78 Proc. IEEE 301 (1990) (summarizing early “high-level language” oriented program synthesis); see also McConnell, supra note 22, at 1 (arguing that the number of software bugs is constant when measured by lines of written code). Specifications can also be informal and take the form of anything from a mental model of a system in the mind of a programmer to a detailed written document describing all goals and use cases for a system. The world of industrial software development is full of paradigms and best practices for producing specifications and building code that meets them.
requirements of public policy. The methods described in this Article are designed to forward, rather than to foreclose, debates about what laws mean and how they ought to work. Our approach aims to empower the policy process, by empowering the policymaker’s tools for dealing with ambiguity and lack of precision, namely review and oversight. We wish to show that software does work as described, allowing a reviewer to determine precisely which properties of the software created a particular rule enforced for a particular decision. Further, if a precise specification of a policy does exist, we wish to show that software which claims to implement that policy in fact does so.

The specification of a system is a critical question for assessment, and system implementers should be prepared to describe the invariants that their system provides. Verification allows the claims of a system’s implementer to constitute evidence that the software in question in fact satisfies those claims.\(^{30}\) Without strong evidence of a computer system’s correctness, even the author of that system cannot reliably claim that it will behave according to a desired policy, and no policymaker or overseer should believe such a claim. For example, a medical radiation device with a software control module was approved for use on patients based on the manufacturer’s claims, but a subtle bug in the software allowed it to administer unsafe levels of radiation, which resulted in six accidents and three deaths.\(^{31}\)

Computer scientists evaluate programs using two testing methodologies: (1) static methods, which look at the code without running the program; and (2) dynamic methods, which run the program and assess the outputs for particular inputs or the state of the program as it is running.\(^{32}\) Dynamic methods can be divided into (1) observational methods in which an analyst can see how the program runs in the field with its natural inputs; and (2) testing methods, which are more powerful, where an analyst chooses inputs and submits them to the program.\(^{33}\)

1. Static Analysis: Review from Source Code Alone

Reading source code does allow an analyst to learn a great deal about how a program works, but it has some major limitations. Code can be complicated or obfuscated, and even expert analysis often misses eventual problems with the behavior of the program.\(^{34}\) For example, the Heartbleed security flaw was a potentially catastrophic vulnerability for most internet users that was caused by a

\(^{30}\) See Carlo Ghezzi et al., Fundamentals of Software Engineering (2d ed. 2002).

\(^{31}\) The commission reviewing the accidents determined that overconfidence on the part of engineers and operators led to both a failure to prevent the problem in the first place and a failure to recognize it as a problem even after multiple accidents had occurred. For an overview, see Nancy Leveson, Medical Devices: The Therac-25, in Safeware: System Safety and Computers app. (1995), an update of the earlier article Nancy G. Leveson & Clark S. Turner, An Investigation of the Therac-25 Accidents, Computer, July 1993, at 18.

\(^{32}\) See Flemming Nielson et al., Principles of Program Analysis (1999).

\(^{33}\) See id. at 1.

\(^{34}\) As an example, the Heartbleed bug was in code that was subjected to significant expert review and careful static analysis with industry leading tools, but was still missed for years. See Edward W. Felton & Joshua A. Kroll, Heartbleed Shows Government Must Lead on Internet Security, 311 Sci. Am. no. 1, 2014, at 14.
common programming error—but that error made it through an open source vetting process and then sat unnoticed for two years, even though anyone was free to read and analyze the code during that time. While there exist automated tools for discovering bugs in source code, even best-of-breed commercial solutions designed to discover exactly this class of error did not find the Heartbleed bug because its structure was subtly different from what automated tools had been designed to recognize. This experience underscores how difficult it can be to find small and simple mistakes. More complex errors evade scrutiny even more easily.

Further, static methods on their own say nothing about how a program interacts with its environment. A program that examines any sort of external data, even the time of day, may have different behavior when run in different contexts. For example, it has long been programming practice to use the time of day as the starting value for a chaotic function used to produce random numbers in programs that do statistical sampling. Such programs naturally choose a different sample of data based on the time of day when they were started, meaning that their output cannot be reproduced unless the time is explicitly represented as an input to the program.

Depending on the technology used to implement a program, static analysis might lead to incomplete or incorrect conclusions simply because it fails to consider the dependencies—that is, the other software that a given program needs in order to operate correctly. For some technologies, the same line of code can have radically different meanings based on the version of even a single dependency. Because of this, it is necessary for static analysis to cover a large portion of any system, and to include at least some dynamic information about how a program will be run.

Within limits, static methods can be very useful in establishing facts about a program, such as the nature of the data it takes in, the kind of output it can

36 Id.
37 See Nielson et al., supra note 32, at 1.
38 This practice dates back at least as far as the 1989 standard for the C programming language. See ANSI X3.159-1989 “Programming Language C.”
40 For example, it is common to re-use “library” code, which provides generic functionality and can be shared across many programs. Ghezzi et al., supra note 30. Library functions can be very different from version to version, meaning that running a program with a different version of the same library can radically change its behavior. It can even change programs that fail to run at all into running, working programs. Because Microsoft Windows refers its system libraries as “Dynamic Linked Libraries,” developers often call this “DLL Hell.” Rick Anderson The End of DLL Hell, Microsoft.com (Jan. 11, 2000). Further, in some programming languages, such as PHP, the meaning of certain statements is configurable. See The Configuration File, PHP, http://php.net/manual/en/configuration.file.php) [https://perma.cc/9LFC-62D9] (describing configuration of PHP).
produce, the general shape of the program, and the technologies involved in the
program’s implementation. In particular, static analysis can reveal the kinds of
inputs that might cause the program to behave in particular ways. Analysts can
use this insight to test the program on different types of inputs. Advanced analysis
can, in some cases, determine aspects of a program’s behavior and establish
program invariants, or facts about the program’s behavior which are true regardless
of what input data the program receives. Programs which are specially designed
to take advantage of more advanced analysis techniques can enable an analyst to
use static methods to prove formally complex invariants about the program’s
behavior. On the simplest level, some programming languages are designed to
prevent certain classes of mistakes. For example, some are designed in such a way
that it is impossible to make the mistake that caused the Heartbleed bug. These
techniques have also been deployed in the aviation industry, for example, to ensure
that the software that provides guidance functionality on rockets, airplanes,
satellites, and scientific probes does not ever crash, as software failures have caused
the losses of several vehicles in the past. More advanced versions of these

41 See Nielson et al., supra note 32, at 1.
42 In programming languages, the most basic structure for expressing behavior that depends on a
value is a conditional statement, often written as if X then Y else Z. A conditional statement will
execute certain code (Y) if the condition (X) is met and different code (Z) if the condition is not
met. Static analysis can reveal where a program has conditional logic, even if it may not always be
able to determine which branch of the conditional logic will actually be executed. For example,
static analysis of conditional logic might show an analyst that a program behaves one way for inputs
less than a threshold and another way otherwise, or that it behaves differently on some particular
special case. Generalizing this analysis can allow analysts to break the inputs of a program into
classes and evaluate how the program behaves on each class. For an overview of logical constructs
in computer programs, see Harold Abelson et al., Structure and Interpretation of Computer Programs
(2d ed. 1996).
43 See Hoare, supra note 25, at 1.
44 See Nielson et al., supra note 32, at 1.
45 Since Heartbleed was caused by improper access to the program’s main memory, see Felten &
Kroll, supra note 35, computer scientists refer to the property that a program has no such improper
access as “memory safety.” For a discussion of the formal meaning of software safety, see Pierce,
supra note 24, at 1. For an approachable description of possible memory safety issues in software,
Several modern programming languages are memory safe, including some, such as Java, that are
widely used in industrial software development. However, while any system could be written in a
memory safe language, developers often choose memory unsafe languages for performance and
other reasons.
46 Both the Ariane 5 and Mars Polar Lander crashed due to software failures. See J.L. Lions,
Chairman, Ariane 501 Inquiry Bd., Ariane 5: Flight 501 Failure (1996); James Gleick, Little Bug,
bug-big-bang.html [https://perma.cc/D4JE-V7K2]; see also Arden Albee et al., JPL Special Review
[https://perma.cc/RE9Z-PX6L]. Similarly, a software configuration error caused the crash of an
Airbus A400M military transport. Sean Gallagher, Airbus Confirms Software Configuration Error
Caused Plane Crash, Ars Technica (June 1, 2015), http://arstechnica.com/information-
techniques may eventually lead to strong invariants being much more commonly and less expensively used in a wider range of applications.

Transparency advocates often claim that by reviewing a program’s disclosed source code, an analyst will be able to determine how a program behaves. Indeed, the very idea that transparency allows outsiders to understand how a system functions is predicated on the usefulness of static analysis. But this claim is belied by the extraordinary difficulty of identifying even genuinely malicious code ("malware"), a task which has spawned a multibillion-dollar industry based largely on the careful review of code samples collected across the internet. Of course, under some circumstances, transparency can also use dynamic methods such as emulating disclosed code on disclosed input data. We discuss transparency further in Part II.

2. Dynamic Testing: Examining a Program’s Actual Behavior

By running a program, dynamic testing can provide insights not available through static source code review. But again, there are limits. While static methods may fail to reveal what a program will do, dynamic methods are limited by the finite number of inputs that can be tested or outputs that can be observed. This is important because decision policies tend to have many more possible inputs than a dynamic analysis can observe or test. Dynamic methods, including structured auditing of possible inputs, can explore only a small subset of those potential inputs. Therefore, no amount of dynamic testing can make an observer certain that he or she knows what the computer would do in some other situation that has yet to be tested.

---


48 Malware analysis can also be dynamic. A common approach is to run the code under examination inside an emulator and then examine whether or not it attempts to modify security-restricted portions of the system’s configuration. For an overview, see Manuel Egele et al., A Survey on Automated Dynamic Malware-Analysis Techniques and Tools, 44 ACM Computing Surveys (CSUR) 6 (2012).

49 Computer scientists call this problem “Combinatorial Explosion.” It is a fundamental problem in computing affecting all but the very simplest programs. Edward Tsang, Combinational Explosion, U. Essex (May 12, 2005), http://cswww.essex.ac.uk/CSP/ComputationalFinanceTeaching/CombinatorialExplosion.html.

50 Even auditing techniques that involve significant automation may not be able to cover the full range of possible input data if that range cannot be limited in advance to a small enough size to be searched effectively. For programmers testing their own software, achieving complete coverage of a program’s behavior by testing alone is considered impossible. Indeed, if testing for the correct behavior of a program were possible at a modest cost, then there would be no bugs in modern software. For a formal version of this argument, see H.G. Rice, Classes of Recursively Enumerable Sets and Their Decision Problems, 74 Transactions Am. Mathematical Soc’y, 358 (1953).

51 Computer security experts often worry about so-called “back doors,” which are unnoticed modifications to software that cause it to behave in unexpected, malicious ways when presented with certain special inputs known only to an attacker. There are even annual contests in which the organizers "propose a challenge to coders to solve a simple data processing problem, but with covert malicious behavior. Examples include miscounting votes, shaving money from financial
Dynamic testing can be divided into “black-box testing,” which considers only the inputs and outputs of a system or component, and “white-box testing,” in which the structure of the system’s internals is used to design test cases. Intuitively, white-box evaluation is more powerful, since any test that can be performed in a black-box setting can also be performed in a white-box setting, but white-box evaluation can suggest more robust test cases by showing an analyst when multiple tests will yield the same behavior or what inputs are likely to trigger differences in the output.52 White-box analysis also helps the developers and operators of a system determine how to monitor its operation so that deviations from expected behavior (e.g., unforeseen bugs, security compromise, abuse, and other unexpected behavior) can be detected and remedied.53

Even structured “audits” of software systems, in which systems are provided with related inputs and analyzed for differential behavior, cannot provide complete coverage of a program’s behavior for the same reason: this methodology explains little about what happens to inputs which have not been tested, even those that differ very slightly.54 Additionally, auditing that treats a system as a black box tells an analyst very little about why differential behavior was observed. A computer program could treat two inputs very differently because it has been explicitly designed to use special case logic for one or both, because those inputs naturally fall into different decision categories, or because the decision rule in use is very sensitive to small changes in its input.

One extremely straightforward and very commonly used form of dynamic program review comes from the practice of logging, or recording certain program actions in a file either immediately before or immediately after they have taken place.55 Analysis of log messages is among the easiest and is perhaps the most common type of functional review performed on most software programs. However, analyzing program logs requires that programs be written to log when they perform events which might be interesting for analysis (and that they log enough information about those events to actually perform the analysis in transactions, or leaking information to an eavesdropper. The main goal, however, is to write source code that easily passes visual inspection by other programmers.” The Underhanded C Contest, http://www.underhanded-c.org/_page_id_2.html [https://perma.cc/82N4]. Back doors have been discovered sitting undetected for many years in commercial, security-focused infrastructure products subject to significant expert review, including the Juniper NetScreen line of devices. See Matthew Green, On the Juniper Backdoor, Few Thoughts on Cryptographic Engineering (Dec. 22, 2015). http://blog.cryptographyengineering.com/2015/12/22/on-juniper-backdoor [https://perma.cc/M7S8-SCM4] (describing the unauthorized code that created a security vulnerability in the Juniper devices).

52 See Glenford J. Myers et al., The Art of Software Testing 1 (3d ed. 2011).
53 See Slawek Ligus, Effective Monitoring and Alerting 1-2 (2012) (describing how to perform monitoring effectively, as opposed to verifying a system’s behavior through testing alone).
54 See infra note 59 and accompanying text.
And because logs are just like other files on a computer, they can easily be modified and rewritten to contain a sequence of events that bears no relation to what a system’s software actually did. Because of this, audit logs meant to record sensitive actions requiring reliable review are generally access controlled or sent to special restricted remote systems dedicated to receiving logging data.

3. The Fundamental Limit of Testing: Noncomputability

Testing of any kind is, however, a fundamentally limited approach to determining whether any fact about a computer system is true or untrue. There are some surprising limitations to the ability to evaluate code statically or dynamically. Counterintuitively, the power of computers is generally limited by a concept that computer scientists call noncomputability. In short, certain types of problems cannot be solved by any computer program in any finite amount of time. There are many examples of noncomputable problems, but the most famous is Alan Turing’s “Halting Problem,” which asks whether a given program will finish running (“halt”) and return an answer on a given input or will run forever on that input. No algorithm can solve this problem for every program and every input. As a corollary, no testing regime can establish any property for all possible programs, since no regime can even determine whether all programs will actually terminate. A related theorem, proposed by Rice, strongly limits the theoretical effectiveness of testing, saying that for any nontrivial property of a program’s behavior, no algorithm can always establish whether a program under analysis has that property. Any such algorithm must get some cases wrong even if the algorithm can do both static and dynamic analyses of the program. However, testing can be very useful in establishing certain specific invariants on restricted classes of programs, and can be made much more useful when programs are designed to facilitate the use of testing to establish those invariants. That is, while testing is not guaranteed to work

57 A common feature of security breaches of computer systems is that attackers will rewrite logs to prevent investigation into how the attack was carried out or who did it. See Dr. Eric Cole et al., Network Security Bible 198 (2005) (noting that the “early stages of an attack often deal with deleting and disabling logging”). Modifying logs in this way can even allow attackers to avoid losing access to a compromised system once the compromise has been detected, since it obscures what steps must be taken to remediate the intrusion. See generally id. (describing how security breaches happen, how they are investigated, and how attackers try to cover the traces of their activity).
60 To see why this is so, imagine writing a new program which halts if it decides that the program it is testing has a certain property, and which runs forever otherwise. For a more detailed version of this argument, see Sipser, supra note 58, at 61.
61 Rice, supra note 50.
62 Id.
in general, it can often be useful in specific cases, especially when those cases have been designed to facilitate testing.

While both static and dynamic methods are after-the-fact assessments--they take the computer system and its design as a given--using both approaches together is often helpful. If an analyst can establish through static methods that a program behaves identically over some class of inputs, the analyst can test a single input from that class and infer the program behavior for the rest of the class. However, not every computer program will be able to be fully analyzed, even with such a combination of methods.

B. The Importance of Randomness

Randomness is essential to the design of many computer systems, so any approach to accountability must grapple with it. However, when randomness is used, it is easy to lose accountability, since by definition any outcome which a randomized process could have produced is at least facially consistent with the design of that process. Accountability for randomized processes must determine why randomness was needed and determine that the source of that randomness and its incorporation into the process under scrutiny meets those goals.

The most intuitive benefit of randomness in a decision policy is that it helps prevent strategic behavior--i.e., “gaming” of a system. When a tax examiner, for example, uses software to choose who is audited, randomization makes it

---

63 This can be done, for example, by noting where in a program’s source code it considers input values and changes its behavior. See infra note 68, at 161 and accompanying text.

64 In fact, there is suggestive theoretical evidence that the power of randomness may be fundamental: there are problems for which the best known randomized algorithm performs much better than the best known deterministic algorithm. For example, the well-studied “multi-armed bandit” problem in statistics has seen wide application in the field of machine learning, where randomized decisionmaking strategies are provably more efficient than nonrandomized ones. See, e.g., J.C. Gittins, Bandit Processes and Dynamic Allocation Indices, 41 J. Royal Stat. Soc’y 148, 148 (1979) (providing a formal mathematical definition of the multi-armed bandit problem); see also Richard S. Sutton & Andrew G. Barto, Reinforcement Learning 1 (1998) (providing a general overview of the usefulness of the multi-armed bandit problem in machine learning applications).

Even outside of machine learning, there are strong indications in computer science theory that certain problems can be solved efficiently only via randomized techniques. Although it is obvious that every efficient algorithm also has an efficient randomized version (which is just rewritten to take some random bits as input and ignore them), it is conjectured but not known that the converse is not true, namely that every efficient randomized algorithm also has a deterministic version that solves the same problem with comparable efficiency. For a summary of work in this area, see Leonid A. Levin, Randomness and Nondeterminism, 1994 Proc. Int’l Congress Mathematicians 1418.

Many important problems, from finding prime numbers (which is necessary for much modern cryptography), to estimating the volume of an object (which is useful in computer graphics and vision algorithms), to most machine learning, had well-understood randomized algorithms that solved them long before they had efficient deterministic solutions (many still do not have any known efficient and deterministic algorithms). For an example, see Morton, infra note 107.

65 For example, a winning lottery ticket with the numbers “1 2 3 4 5” is just as likely to be correct as any other ticket, and yet it seems strikingly unlikely. In a similar way, it will always be necessary when randomness is involved in a process to ensure that even outcomes that are “correct” in the sense that the system could have produced them are also correct in the sense that they fulfill the goals which necessitated randomness in the first place (e.g., in a lottery, that the winning ticket numbers not be known in advance of their selection and not be influenced by the lottery operators).
impossible for a taxpayer to be sure whether or not he or she will be audited. Those who are evading taxes, in particular, face an unknown risk of detection—which can be minimized, but not eliminated—and do not know whether, or when, they should prepare to be audited. Similarly, if additional security screening is applied at random to those crossing a checkpoint, or if the procedures at the checkpoint are changed at random on a day-to-day basis, a smuggler or attacker cannot be as prepared as if the procedures were fully known in advance.66 Additionally, studies of the performance of human guards have shown that randomization in procedures reduces boredom, thereby improving vigilance.67

The card game of poker illustrates a second benefit: randomness can obscure secret information. A good poker player has secret information—how good her cards are—that affects how she will bet. By occasionally bluffing, she randomizes her behavior and makes it more difficult for opponents to infer the quality of her hand.

In situations where a scarce or limited resource must be apportioned to equally deserving recipients such that not all qualified applicants can receive a share, randomness can help by fairly apportioning resources to participants in a way that cannot be controlled or predicted by those in control of the resource. For example, the Diversity Visa Lottery, considered in Part II, is a case where a random lottery allocates a scarce resource—a limited number of visas to live and work in the United States. Randomness as a source of fairness requires two attributes: first, the outcome must not be controlled by the system’s operator, or else the randomness serves little purpose when compared to a model where the system’s operator just chooses the winners; and second, the outcome must not be predictable, or else the operator of such a system could put its favored winners into certain slots or slip them “winning tickets” prior to the system’s operation. Further, it is important that the random choices made when the system is run be binding upon the system’s operator, so that the system cannot be run many times to control the eventual output by shopping for a favorable result among many actual runs of the system. We explore precisely how to address these issues below.

Many machine learning systems use randomization as part of their normal operation. It turns out that guessing randomly and adjusting the probability of each class of output often leads to much better performance than trying to determine the absolute best decision at any point.68

66 See, e.g., James Pita et al., Deployed ARMOR Protection: The Application of a Game Theoretic Model for Security at the Los Angeles International Airport, 2008 Proc. 7th Int’l Conf. on Autonomous Agents & Multiagent Sys.: Industry & Applications Track 125 (describing a software system that uses a game-theoretic randomized model to improve the efficiency of police and federal air marshal patrols at the Los Angeles International Airport).
67 See, e.g., Richard L. Thackray et al., FAA Civ. Aeromedical Institute, Physiological, Subjective, and Performance Correlates of Reported Boredom and Monotony While Performing a Simulated Radar Control Task (1975) (discussing the improvement of performance through increased unpredictability in procedures).
Finally, randomization can give computers more flexibility to perform well in unexpected environments. Consider how the Roomba robot is programmed to vacuum rooms.\textsuperscript{69} If rules of motion were hard-coded in the software controlling the robot, an unusual furniture configuration might lead to the Roomba getting stuck in a corner or under a table or repeatedly following the same path without cleaning the rest of the room. Adding in randomized motion allows it to escape these patterns and work more effectively without the need to code in all possible room configurations. By allowing for unknowns, randomized strategies can avoid worst-case outcomes with high probability, no matter how unfriendly the environment turns out to be.\textsuperscript{70}

However, poorly designed randomization can lead to unaccountable automated decisions. If a decision depends on a randomly selected value, then any outcome consistent with some possible value of the random choice, no matter how unlikely, must be considered valid. Concretely, if a decision is based on the outcome of a coin flip, even if the coin is biased to land heads up 99 times out of 100, a result based on a tails up flip cannot be shown to be improper, since one out of every 100 results will be derived from the value of tails. A corrupt decisionmaker could influence this supposedly random choice, picking the value of the coin consistent with its preferred outcome, or could flip many coins and then assign the value of each flip to the set of decisions he has to make (perhaps by changing the order in which he considers decisions) in such a way as to pick the outcomes he desires. Random choices generated by a computer system can also be remade by re-running a program until the outputs match a preferred outcome. Without designing the computer system to demonstrate that this is not happening, it is very hard for a decisionmaker to prove that he has not done this. The speed of automated decisionmaking only increases this risk; while physical randomization of balls in a tumbler can only produce a small number of values per hour of effort, a computer can try thousands or millions of outcomes in a matter of minutes.

\textsuperscript{69} For a description of the Roomba’s movement algorithm, see Ja-Young Sung et al., “My Roomba is Rambo”: Intimate Home Appliances, in 2007 Proc. 9th Int’l Conf. on Ubiquitous Computing 145.

\textsuperscript{70} More concretely, one study showed that computer-generated teaching plans customized to particular students can be less effective than lesson plans without customization if the software model used to tailor lessons to individual performance is trained on large groups that do not capture individual-specific patterns. This failure of “big data” methods trained on large groups of students to properly capture the quirks of a “small data” situation (such as a classroom-sized group of students) can be avoided by adding random deviations from the model’s prediction and tracking the results of these deviations. See, e.g., Yun-En Liu et al., Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits, 2014 Proc. 7th Int’l Conf. on Educ. Data Mining 161 (observing the introduction of small changes on result prediction).
Additionally, a randomized process is not easily reproduced. For example, if it depends on interaction with its environment (e.g., the operating system on which it is running, its human user, or a database with rapidly changing records), its behavior may be altered in a nondeterministic way since that environment can change between runs.\textsuperscript{71} One unwieldy solution to this problem is to capture all of the environment in which a program runs, so that this environment can be replayed by an analyst. However, this solution does not address how to verify all of the reasons that randomness might be needed in a process.

II. Designing Computer Systems for Procedural Regularity

The first goal in any plan to govern automated decisionmaking should be to enable the people overseeing the process—whether they are government officials, corporate executives, or members of the public—to know how a computer system makes decisions (or, at the very least, that it makes decisions based on some rule, even if that rule is not fully disclosed). A baseline requirement in most contexts is procedural regularity: each participant will know that the same procedure was applied to her and that the procedure was not designed in a way that disadvantages her specifically.\textsuperscript{72} This baseline requirement draws on the Fourteenth Amendment principle of procedural due process. Ever since a seminal nineteenth century case, the Supreme Court has articulated that procedural due process requires rules to be generally applicable and not designed for individual cases.\textsuperscript{73} Similarly, federal statutes articulate the requirement for procedural regularity in administrative agency actions.\textsuperscript{74} These principles are also enshrined in the Federal Rules of Civil Procedure for civil litigation.\textsuperscript{75}

In this Part, we will demonstrate that the tools of computer science can guarantee important elements of procedural regularity when they are incorporated in the initial design of computer systems. Specifically, these tools can assure that:

- The same policy or rule was used to render each decision.
- The decision policy was fully specified (and this choice of policy was recorded reliably) before the particular decision subjects were known, reducing the ability to design the process to disadvantage a particular individual.

\textsuperscript{71} One specific example is a program that chooses a random value based on the time that it has been running but takes different amounts of time to run based on what other programs are running on the same physical computer system.

\textsuperscript{72} For example, a tax auditing risk assessment should not single out individuals either by name or by identifying characteristics. If a process added extra weight to filers of a particular postal code, gender, and birth month, this could be enough to single out individuals in many cases. See, e.g., Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. Rev. 1701, 1716-27 (2010) (showing that an individual’s identity may be reverse-engineered from a small number of data points).

\textsuperscript{73} See Marchant v. Pa. R.R., 153 U.S. 380, 386 (1894) (holding that the plaintiff had due process because “her rights were measured, not by laws made to affect her individually, but by general provisions of law applicable to all those in like condition”).

\textsuperscript{74} See Administrative Procedure Act, 5 U.S.C. §§ 551-59 (2012) (prescribing exhaustive procedural requirements for most levels of federal administrative agency action).

\textsuperscript{75} See Fed. R. Civ. P. 1 (noting that the rules apply “in all civil actions and proceedings . . . to secure the just . . . determination of every action and proceeding” (emphasis added)).
• Each decision is reproducible from the specified decision policy and the inputs for that decision.
• If a decision requires any randomly chosen inputs, those inputs are beyond the control of any interested party.

After describing these properties and showing how they can be implemented, we will apply them to a case study—the Diversity Visa Lottery at the State Department—where application of these tools could greatly improve the legitimacy and fairness of an automated decision procedure.

A. Transparency and Its Limits

A naïve solution to the problem of verifying procedural regularity is to demand transparency of the source code as well as inputs and outputs for the relevant decisions; if all of these elements are public, it seems easy to determine whether procedural regularity was satisfied. Indeed, full or partial transparency can be a helpful tool for governance in many cases, and transparency has often been suggested as a remedy to accountability issues for computerized systems. However, transparency alone is not sufficient to provide accountability in all cases.

First and foremost, it is often necessary to keep secret the elements of a decision policy, the computer systems that implement it, key inputs, or the outcome. Keeping aspects of a decision policy secret can help prevent strategic “gaming” of a system. For example, the IRS may look for signs in tax returns that are highly correlated with tax evasion based on returns previously audited. But if the public knows exactly which things on a tax return are treated as telltale signs of fraud, tax cheats may adjust their behavior and the signs may lose their predictive value for the agency. Moreover, when the decision being regulated is a commercial one, such as a decision to offer credit, a business’s legitimate interest in protecting proprietary information or guarding trade secrets like the underwriting formula may be incompatible with full transparency. And in many contexts, an automated decision may use as inputs, or will create as an output, sensitive or private data that should not be broadly shared to protect business interests, privacy, or the integrity of law enforcement or investigative methods. In some cases, especially with financial or medical data, disclosure may be barred or limited by statutes or regulations. Finally, in many situations—such as scoring consumers for credit or insurance risk—the purpose of the automated decision process is to determine something not

78 See Jeff Reeves, IRS Red Flags: How to Avoid a Tax Audit, USA Today (Mar. 15, 2015 12:08 PM), http://www.usatoday.com/story/money/personalfinance/2014/03/15/irs-tax-audit/5864023 [https://perma.cc/BFW5-DG34] (identifying characteristics of tax returns that trigger IRS audit).
directly measurable, such as the risk of defaulting on credit or claiming a loss on an insurance policy. Because these values cannot be measured directly, they are computed from proxy variables such as a consumer’s credit history, income, or personal attributes. Consumers who understand these actuarial processes would be rational in attempting to control the input proxy variables, which in turn could render the scoring process less useful in elucidating unmeasurable risk. See supra Part I. Secrecy discourages strategic behavior by participants in the system and prevents violations of legal restrictions on disclosure of data.

Second, while transparency allows for the testing strategies described earlier (i.e., static and dynamic tests including auditing), those methods are often insufficient to verify properties of software systems if these systems have not been designed with future evaluation and accountability in mind. See generally Cathy O’Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016). In economic policymaking, this is sometimes known as Goodhart’s Law, popularly formulated as “[w]hen a measure becomes a target, it ceases to be a good measure;” Goodhart formulated it more formally as “[a]ny observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” C.A.E. Goodhart, Problems of Monetary Management: The U.K. Experience, in 1 Papers in Monetary Economics (1976). Hardt and his co-authors have developed adversarial methods for designing automated decision and classification systems that remain robust even in the face of gaming. See Moritz Hardt et al., Strategic Classification, 2016 Proc. 2016 ACM Conf. on Innovations Theoretical Computer Sci. 111 (discussing methods to strengthen classification models).

Third, for decision processes that involve some element of randomness, even full transparency--of the system’s source code, its inputs, its operating environment, and its results--does not foreclose the possibility that an outcome could be improperly fixed in an undetectable way, as described in Section I.C. A classic lottery provides an excellent example. A perfectly transparent algorithm that uses a random number generator to assign a number to each participant and has the participants with the lowest numbers win will yield results that cannot be reproduced or verified because the random number generator will produce different random numbers each time. Reviewing the code alone, or even together with the data fed into it and the environment in which it was operated, does not tell us that it was actually used to produce a particular result. By design, the process produces unpredictable results that are not reproducible.

Fourth and finally, systems that change over time cannot be fully understood through transparency alone. System designers regularly change complicated automated decision processes--such as search engine ranking

---

80 In particular, consumers are rational to modify proxy variables that control their perceived risk when those variables are cheaper or easier to manipulate than the gain obtained via better treatment by the decision system. Intuitively, if proxy variables are weak and easy to alter or sometimes poorly correlated with the feature being measured (e.g., standardized test scores as a measure of student learning), they are more likely to be gamed than features which are highly proximate to the value being estimated, or which are difficult or expensive to alter (e.g., annual income as a measure of creditworthiness in a particular transaction). See generally Cathy O’Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016). In economic policymaking, this is sometimes known as Goodhart’s Law, popularly formulated as “[w]hen a measure becomes a target, it ceases to be a good measure;” Goodhart formulated it more formally as “[a]ny observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” C.A.E. Goodhart, Problems of Monetary Management: The U.K. Experience, in 1 Papers in Monetary Economics (1976). Hardt and his co-authors have developed adversarial methods for designing automated decision and classification systems that remain robust even in the face of gaming. See Moritz Hardt et al., Strategic Classification, 2016 Proc. 2016 ACM Conf. on Innovations Theoretical Computer Sci. 111 (discussing methods to strengthen classification models).

81 See supra Part I.

82 There are ways to incorporate randomness that can be replicated. See infra subsection II.B.3.
methodology, spam filter rules, intrusion detection system methods, or the algorithms that select website ads— in response to strategic behavior by participants in the system. Computer systems that choose social media posts to display to users might respond to user behavior. “Online” machine learning systems can update their model for predictions after each decision, incorporating each new observation as part of their training data. Even knowing the source code and data for such systems is not enough to replicate or predict their behavior—we also must know precisely how and when they interacted or will interact with their environment. Whether updates to the system are effected by human engineers and operators (e.g., a search engine engineer modifies the software used to rank pages) or by a machine learning system (e.g. the search engine’s software discovers that users more often click the second link for a certain query instead of the first, so it reverses their order), transparency alone does little to explain either why any particular decision was made or how fairly the system operates across bases of users or classes of queries. With such systems, there is the added risk that the rule disclosed is obsolete by the time it can be analyzed. Online machine learning systems update their decision rules after every query, meaning that any disclosure will be obsolete as soon as it is made.

B. Auditing and Its Limits

Beyond transparency, auditing is another strategy for verifying how a computer system works. An audit treats the decision process as a black box, whose inputs and outputs are visible but whose inner workings are unseen. The approach has a long history in offline contexts, such as testing for discrimination in retail car negotiations. For retail car negotiations, the transparency of the bargaining

84 Many spam filters work by keeping a list of bad terms, email addresses, and computer to block messages from. The most widely used “blacklist” is produced by the organization Spamhaus. See SBL Advisory, Spamhaus, https://www.spamhaus.org/sbl [https://perma.cc/V9LN-EPK9] (describing the Spamhaus Block List Advisory, “a database of IP addresses from which Spamhaus does not recommend the acceptance of electronic mail”).
86 See Jure Leskovec et al., Mining of Massive Data Sets ch. 8 (2d ed. 2014).
88 See Ian Ayres, Fair Driving: Gender and Race Discrimination in Retail Car Negotiations, 104 Harv. L. Rev. 817, 818 (1991) (using auditing to determine “[w]ether the process of negotiating for a new car disadvantages women and minorities”).
process for the purchase of a car is insufficient to determine if different prices are offered based on race or gender.\textsuperscript{89}

In computer science, auditing refers to the review of digital records describing what a computer did in response to the inputs it received.\textsuperscript{90} Auditing is intended to reveal whether the appropriate procedures were followed and to uncover any tampering with a computer system’s operation. For example, there is a substantial body of literature in computer science that addresses audits of electronic voting systems,\textsuperscript{91} and security experts generally agree that proper auditing is necessary but insufficient to assure secure computer-aided voting systems.

Computer scientists, however, have shown that black-box evaluation of systems is the least powerful of a set of available methods for understanding and verifying system behavior.\textsuperscript{92} Even for measuring demonstrable properties of software systems, such as testing whether a system functions as desired without bugs, it is much more powerful to be able to understand the design of that system and test it in smaller, simpler pieces.\textsuperscript{93} Approaches that attempt to review system failures simply by looking at how the output responds to changes in input are limited by either an inability to attribute a cause to those changes or an inability to interpret whether or why a change is significant.\textsuperscript{94} Instead, software developers regularly use other, more powerful evaluation techniques.\textsuperscript{95} These include white-box testing (in which a the person doing a test can see the system’s code, and uses that knowledge to more effectively search for bugs) and using programming languages that automatically preclude certain types of mistakes.\textsuperscript{96}

\textsuperscript{89} See id. at 827-28 (observing that women and minorities received worse prices than white males even when using the same negotiation strategy).
\textsuperscript{92} Specifically, white-box testing, in which an analyst has access to the source code under test, is generally considered to be superior; even in cases where the basic testing approach does not make use of the structure of the software (e.g., so-called “fuzz testing” where a program is subjected to randomly generated input), testing benefits from some access to the structure of programs. See supra note 52 and accompanying text.
\textsuperscript{93} Also, consider the difficulties encountered in one such audit study. The authors show a causal relationship between changing sensitive, protected attributes (e.g., gender) and the advertisements presented to a user (e.g., advertisements for high-paying jobs). See Amit Datta et al., Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination, 2015 Proc. Privacy Enhancing Techs. 92, 105-06. However, the methodology is unable to identify the mechanism of this causation or even whether the results discovered will generalize beyond the data seen in the study. Id. at 105.
\textsuperscript{94} For example, if the output of a system is an error or other failure such as a crash, it is not obvious to an analyst how to modify the output to learn much at all.
\textsuperscript{95} See supra Section I.A.
\textsuperscript{96} See supra note 24.
C. Technical Tools for Procedural Regularity

As we demonstrated above, transparency and auditing often do not suffice for accountability. In this Section, we introduce computational methods that can provide accountability for procedural regularity even when some information is kept secret. These methods can be used alongside transparency and auditing when appropriate and apply to all computer systems.97

Our approach harnesses the power of computational methods and does not take the design of the computer system as a given. Instead, we explicitly advocate for systems to be designed to use methods such as the ones described here. Nor do we give up on governance when all or part of a computer system must remain secret. We rely on several advanced techniques from computer science to enable the governance of secret decision systems: software verification, cryptographic commitments, zero-knowledge proofs, and fair random choices. Counterintuitively, even when a piece of software or the data input to it is secret, these methods can guarantee that the software and inputs satisfy the requirements for procedural regularity. They can verify that the same decision policy was used for each decision, that the policy was determined and recorded before inputs were known, and that the outcomes are reproducible. Just because a given computer system or piece of software is secret does not mean that nothing about that system can be known.

1. Software Verification

Software verification is a set of techniques for proving mathematically that software has certain properties, either by analyzing existing code or by building software using specialized tools for extracting proved correct invariants. It has been a promising field for many decades, and while many benefits are only realized in research prototypes today, these methods are finding increasing industrial adoption, especially in sectors where software is safety or security-critical and in domains with strong liability regimes.98 While the complete verification of any program is

97 While the methods we propose are general, they can be inefficient for certain applications. The cost of providing a certain level of accountability must be considered as part of the design of any policy requirement. For more detail, see Kroll, infra note 119.

98 While software verification has been embraced by the aviation and industrial control sectors and for some financial applications (for example, the hedge fund Jane Street regularly touts its use of formal software analysis in recruiting materials sent to computer science students), it has yet to see much adoption in the critical fields of healthcare and automotive control. See Jean Souyris et al., Formal Verification of Avionics Software Products, 2009 Proc. of the 2nd World Congress on Formal Methods 532 (describing the use of software verification at Airbus); Norbert Volker & Bernd J. Kramer, Automated Verification of Function Block-Based Industrial Control Systems, 42 Sci. Computer Programming 101 (2002). Indeed, researchers have effected compromises of embedded healthcare devices such as pacemakers, See, e.g., Daniel Halperin et al., Pacemakers and Implantable Cardiac Defibrillators: Software Radio Attacks and Zero-Power Defenses, 2008 IEEE Symp. on Security and Privacy 129, 141 (finding that implantable cardioverter defibrillators are “potentially susceptible to malicious attacks that violate the privacy of patient information” and “may experience malicious alteration to the integrity of information or state”). News reports also indicate that former Vice President Dick Cheney had the remote software update functionality on his pacemaker disabled so that updating the software would require surgery, ostensibly in order to prevent remote compromise of his life-critical implant. Andrea Peterson, Yes, Terrorists Could Have Hacked Dick Cheney’s Heart, Wash. Post (Oct. 21, 2013),
an expensive undertaking largely reserved to technologists versed in this particular area, it is important as a matter of policy to understand the options that are available so that costs and benefits can be weighed and acted upon.

Unlike static analysis, which aims to examine already-written code for bugs or deviations from its specification, or software testing, which aims to verify that software meets a specific set of functional requirements by explicitly executing the software in a particular configuration, software verification aims to prove invariants about a program mathematically, using logic to reason about a program’s behavior under all conditions.99 Verified programs come with a mathematically checkable proof demonstrating that they have certain invariants, rendering testing for those invariants unnecessary, as the proof implies that such tests will always pass.100

There are many ways to verify software. For instance, a program can be carefully annotated using formal logic to determine its behavior in a precise manner, though this can be expensive and will not always yield the desired invariants;101 a program can be certified by another program which translates it to


Additionally, researchers have also demonstrated spectacular compromises of automobile control systems, including disabling brakes, controlling steering and acceleration, and completely cutting engine power. See Karl Koscher et al., Experimental Security Analysis of a Modern Automobile, 2010 IEEE Symp. on Security and Privacy 447 (performing early analyses of the security of automobile computers); see also Stephen Checkoway et al., Comprehensive Experimental Analyses of Automotive Attack Surfaces 2011 Proc. 20th USENIX Conf. on Security 77 (same). Subsequently, researchers have demonstrated problems in other models, including luxury models with significant telematics capabilities and remote software upgrade capability, showing that active maintenance of these software systems does not completely defend against attacks. See, e.g., Jonathan M. Gitlin, Man Hacks Tesla Firmware, Finds New Model, Has Car Remotely Downgraded, Ars Technica (Mar. 8, 2016 11:36 AM), http://arstechnica.com/cars/2016/03/man-hacks-tesla-firmware-finds-new-model-has-car-remotely-downgraded [https://perma.cc/R9C5-9RTY] (describing an incident where a Tesla car model was hacked despite frequent software updates). Problems with spontaneous acceleration in many Toyota vehicles were later traced to software issues. See Phil Koopman, A Case Study of Toyota Intended Acceleration and Software Safety (Sept. 18, 2014), https://users.ece.cmu.edu/~koopman/pubs/koopman14_toyota_ua_slides.pdf [https://perma.cc/VP9T-VYMF] (presenting a detailed analysis of the issue). And of course, Volkswagen designed its engine control software to defeat an emissions testing regime. For a complete timeline of the Environmental Protection Agency’s actions on this matter, see Volkswagen Light Duty Diesel Vehicle Violations for Model Years 2009-2016, EPA.gov, https://www.epa.gov/vw [https://perma.cc/C83U-UZLG] (last updated Nov. 7, 2016).

99 See supra Section I.A.
100 See supra notes XX-XX.
101 For one of the earliest approaches to representing programs as statements in formal logic, see C.A.R. Hoare, An Axiomatic Basis for Computer Programming, 12 Comm. ACM 576, 576-80 (1969). While Hoare’s techniques form the basis of many modern methods, some methods attempt to build software that is correct by virtue of its construction, rather than analyzing software that has already been written. For an overview of different approaches and their tradeoffs, see B. Bérard et al., Systems and Software Verification: Model-Checking Techniques and Tools (2001). For a classic reference on how to include these techniques in the software engineering process, see Carlo Ghezzi et al., Fundamentals of Software Engineering (2d ed. 2003).
a form which is guaranteed to have the desired property; a program can be exhaustively tested for all possible inputs to ensure that an invariant is never violated; or a program can be built using tools that allow for the careful specification of invariants (and proofs of those invariants). Researchers have even verified entire operating systems using these techniques. Verification can be communicated to clients in a number of ways: so called proof-carrying code comes with a machine-checkable proof of its verified invariants which can be checked by a user just prior to running the program; a user can recompute the analysis used to generate the proof; or the truth of the proof can be confirmed by an entity the user trusts, with cryptography used to guarantee that the version a user is running matches the version that was verified.

102 These tools are known as “certifying compilers.” The advantage of a certifying compiler is that one need only expend effort verifying the certifying compiler itself, not the software being compiled, in order to prove that the desired invariant holds for the compiled software. For a description of the original concept and a first implementation, see George C. Necula & Peter Lee, The Design and Implementation of a Certifying Compiler, 33 ACM SIGPLAN Notices 333 (1998). There are many examples of certified compiler systems. See, e.g., Joshua A. Kroll et al., Portable Software Fault Isolation, 2014 Proc. IEEE 27th Computer Security Found. Symp. 18 (describing a certifying software fault isolation compiler built out of CompCert’s certified back end).

103 This technique, known as “model checking,” could also be described as a form of static analysis. Model checking aims to verify an invariant by finding a counterexample (an input to the program which makes the invariant untrue and hence not an invariant). If a counterexample can be found, the program has a demonstrable bug. If no counterexample can be found, that invariant has been verified. See supra note XX and infra note XX.

104 Several such programming languages exist, though one of the more successful toolkits in active research is the proof assistant Coq, which allows for writing complex programs and theorems and invariants about those programs, in such a way that the proved-correct programs can be “extracted” into executable code. For an introduction to Coq, see Adam Chlipala, Certified Programming with Dependent Types: A Pragmatic Introduction to the Coq Proof Assistant (2013) and Yves Bertot & Pierre Castéran, Interactive Theorem Proving and Program Development: Coq’Art: The Calculus of Inductive Constructions (2004) (describing advanced programming techniques). Several large and complex programs have been written in Coq, which demonstrates that it is a robust tool capable of supporting nontrivial development tasks and proofs of correctness about those tasks. Perhaps the most famous of these was the proof of the “four-color theorem,” which states that any map can be drawn using only four colors such that no border on the map uses the same color for the regions on both sides of the border. Georges Gonthier, Formal Proof--The Four-Color Theorem, 55 Notices AMS 1382 (2008). Similar tools include a theorem prover for programs written in ANSI Common Lisp 2 and the interactive theorem prover Isabelle. See Lawrence C. Paulson, The Foundation of a Generic Theorem Prover, 5 J. Automated Reasoning 363 (describing the design and implementation of Isabelle).


107 This approach would consist of the certifying authority making a cryptographically signed statement that it had verified the proof for a binary with a certain cryptographic hash value and the
However, just because a program has been verified or proven correct does not mean that it has been vetted at all for correctness or compliance with a policy. Verification typically constitutes a proof that the software object in use matches its specification, but this analysis says nothing about whether the specification is sufficiently detailed, correct, lawful, or socially acceptable, or constitutes good policy. Software verification is a rapidly developing field, and the costs of building fully verified software will likely drop precipitously in the coming decades, leading to wide adoption in the software industry due to the benefits of reduced security exposure and the elimination of many types of software bugs.

2. Cryptographic Commitments

A cryptographic commitment is the digital equivalent of a sealed document held by a third party or in a safe place. It is possible to compute a commitment for any digital object (e.g., a file, a document, the contents of a search engine’s index at a particular time, or any string of bytes). Commitments are a kind of promise that binds the committer to a specific value for the object being committed to (i.e., the object inside the envelope) such that the object can later be revealed and anyone can verify that the commitment corresponds to that digital object.\(^\text{108}\) In this way, as in the envelope analogy, an observer can be certain that the object was not changed since the commitment was issued and that the committer did indeed know the value of the object at the time the commitment was made (e.g., the source code to a program or the contents of a document or computer file). Importantly, secure cryptographic commitments are also hiding, meaning that knowledge of the commitment (or possession of the envelope in the analogy) does not confer information about the contents. This gives rise to the sealed document analogy: once an object is “inside” the sealed envelope, an observer cannot see it nor can anyone change it. However, unlike physical envelopes, commitments can be published, transmitted, copied, and shared at very low cost and do not need to be guarded to prevent tampering. In practice, cryptographic commitments are much smaller than the digital objects they represent.\(^\text{109}\) Because of this, commitments can be used to lock in knowledge of a secret (say, an undisclosed decision policy) at a certain time (say, by publishing it or sending it to an oversight body) without revealing the contents of the secret, while still allowing the secret to be disclosed later (e.g., in a court case under a discovery order) and guaranteeing that the secret


\(^{109}\) See id. at 1 (noting that commitments can be smaller than the statements to which they relate). A typical commitment will be 128 or 256 bytes, regardless of the size of the committed object. See Info. Tech. Lab., Nat’l Inst. of Standards & Tech., FIPS PUB 180-4, Secure Hash Standard(2015) (describing the hash algorithms accepted for government computer applications, which provide widely-used standards in industry).
was not changed in the interim (for example, that the decision policy was not modified from one that was explicitly discriminatory to one that was neutral).\textsuperscript{110}

When a commitment is computed from a digital object, the commitment also yields an opening key, which can be used to verify the commitment.\textsuperscript{111} Importantly, a commitment can only be verified using the precise digital object and opening key related to its computation; it is computationally implausible for anyone to discover either another digital object or another opening key which will allow the commitment to verify properly. In the envelope metaphor, this is tantamount to proof that neither the envelope nor the document inside the envelope was replaced clandestinely with a different envelope or document. Any digital object (e.g., a file, document, or any string of bytes) can have a commitment and an opening key such that it is: 1) impossible to deduce the original object from the commitment alone; 2) possible to verify, given the opening key, that the original object corresponds to the commitment, and 3) impossible to generate a fake object and fake opening key such that using the (real) commitment and the fake opening key will reveal the fake object.

Cryptographic commitments have useful implications for procedural regularity in automated decisions. They can be used to ensure that the same decision policy was used for each of many decisions. They can ensure that rules implemented in software were fully determined at a specific moment in time. This means a government agency or other organization can commit to the assertions that (1) the particular decision policy was used, and (2) the particular data were used as input to the decision policy (or that a particular outcome from the policy was computed from the input data). The agency can prove the assertions by taking its secret source code, the private input data, and the private computed decision outcome and computing a commitment and opening key (or a separate commitment and opening key for each policy version, input, or decision). The company or agency making an automated decision would then publish the commitment or commitments publicly and in a way that establishes a reliable publication date, perhaps in a venue such as a newspaper or the Federal Register. Later, the agency could prove that it had the source code, input data, or computed results at the time of commitment by revealing the source code and the opening key to an oversight

\textsuperscript{110} As a curiosity, we remark that the popular board game Diplomacy is essentially based on physical world commitments: each player negotiates a set of moves for the next round of the game, but then these moves are written on paper and passed secretly to a game master who stores them in an envelope. Once all players have entered their moves, the moves are revealed and taken simultaneously. This commitment mechanism allows players to simulate simultaneous moves without any risk that a player will fall behind or change their moves in a particular round in response to their perception of what another player is doing in that round. However, the commitment mechanism alone does not prevent players from entering incorrect or impossible moves, writing nonsense on their paper instead of moves, or simply refusing to enter a move at all (the game master, however, enforces that all moves placed into the envelope are correct and all players must trust her to do this to ensure that the game is not spoiled). Below, in the section on zero-knowledge proofs, we describe how techniques from computer science can address the role of the game master purely through computation without the need for an entity trusted by all players of the game.

\textsuperscript{111} See Oded Goldreich, Foundations of Cryptography – A Primer (2005).
body such as a court. This technique assures that the software implementing the decision policy was determined and recorded prior to the publication of the commitment, which can be useful in demonstrating that neither the software nor the decision policy were influenced by later information or events.

By themselves, however, cryptographic commitments do not prevent the committer from lying and generating a fake commitment that it cannot open at all or from destroying (or refusing to disclose) the information that allows a valid commitment to be opened. In either case, when the time comes to reveal the contents of the commitment, it will be demonstrable that the committer has misbehaved. However, an observer does not know the nature of the misbehavior. The committer may not have a correct opening key (analogous to having sealed an unintelligible or irrelevant document in a physical envelope) or may want to lie about what was in the original file (analogous to discovering that the contents of the envelope may be embarrassing under scrutiny of oversight). In either case, an oversight authority might punish the committer for lying and assume the worst about the contents of the missing file. However, it would be preferable to be able to avoid this scenario altogether, which we can do with another tool, zero-knowledge proofs, described below.

3. Zero-Knowledge Proofs

A zero-knowledge proof is a cryptographic tool that allows a decisionmaker, as part of a cryptographic commitment, to prove that the decision policy that was actually used (or the particular decision reached in a certain case) has a certain property, but without having to reveal either how that property is known or what the decision policy actually is.

For example, consider how money flows in an escrow transaction. Traditionally, an escrow agent holds payment until certain conditions are met. Once they are, the agent attests to this fact and disburses the money according to a predetermined schedule. Zero-knowledge proofs can allow escrow without a trusted agent. Suppose that an independent sales contractor wishes to certify that she has remitted appropriate taxes from her sales in order to be paid by a counterparty, but without revealing precisely how much she was able to sell an item for. Using a zero-knowledge proof, she can demonstrate that sufficient taxes were paid without disclosing her sales prices or earnings to a third party.

Another classic example used in teaching cryptography posits that two millionaires are out to lunch and they agree that the richer of them should pay the bill. However, neither is willing to disclose the amount of her wealth to the other. A zero-knowledge proof allows them both to learn who is wealthier (and thus who should pay the restaurant tab) without revealing how much either is worth.

112 A parallel to this assumption is a spoliation inference, which sanctions a party who withholds, tampers, or destroys evidence by assuming that the missing or changed evidence was unfavorable to the spoliator. See Fed. R. Civ. P. 37(e)(2)(A) (providing that if electronically stored information is lost because a party, intending to deprive the other party of the information, failed to take reasonable steps to preserve it, the court may “presume that the lost information was unfavorable to the party”).

113 See Goldreich, supra note 111, at .
A zero-knowledge proof works with cryptographic commitments to verify procedural regularity in the following manner. If a decisionmaker makes a trio of commitments, $A$, $B$, and $C$, where $A$ is a commitment to the decision policy, $B$ a commitment to the inputs that were used in a particular case, and $C$ a commitment to the decision actually reached in that case, then zero-knowledge proofs let the public verify that $A$, $B$, and $C$ really do correspond to each other. In other words, the decisionmaker can prove that, when the committed policy $A$ is applied to the committed input data $B$, the result is the committed outcome $C$.

This allows decisionmakers to build audit logs, which can be verified by the public to confirm that the decisionmaker applied the appropriate policy to the correct input in order to reach the stated outcome, all without revealing the decision policy itself and without revealing private data that might be included in the input or outcome.

Later, if the outcome is challenged, a court or other oversight body can compel the decisionmaker to reveal the actual policy and input used and can verify that it matches the published commitment, effectively providing digital evidence that the decisionmaker was honest about its announced decision. By using a commitment to the same policy in decisions for multiple decision subjects, a decisionmaker can demonstrate that it is applying a consistent policy across those subjects. Such zero-knowledge proofs can be enhanced to test parts of the decision policy, either by exhibiting properties of the input-output relation (e.g., that a credit score would have been the same if the subject’s gender were reversed) or properties of the policy itself (e.g., that the policy only uses certain inputs for certain purposes).

4. Fair Random Choices

Where random choices are part of a decisionmaking process, the fairness of the randomness used in those computer systems should be verifiable. Poorly designed randomization can lead to unaccountable automated decisions. The decisionmaker could influence the supposedly random choices or could generate many sets of random values and then pick the set that gives its preferred outcome. Additionally, a randomized process is not easily reproduced. For example, if it depends on interaction with its environment (e.g., the operating system on which it is running or its human user), its behavior may be altered in a nondeterministic way since that environment can change between runs.\(^{114}\)

Automated decision processes must therefore be designed from the beginning to allow for oversight of the decisionmaker and to avoid problems with unpredictable behavior. To solve this problem, a decisionmaker could demonstrate that any unpredictable behavior or random choices in the software does not affect the eventual output; for example, a program designed to find the top of a hill (i.e.,

\(^{114}\) One example is a program that chooses a random value based on the time that it has been running but takes different amounts of time to run based on other programs that are running on the same physical computer system.
optimize some objective) can start at any randomly chosen point and take any arbitrary path upwards and will still ultimately return the same maximum value.\textsuperscript{115}

More often, the random choices made by an automated decision process will affect the results. In these cases, the software implementing the decision can always be redesigned to replace the set of random choices made by the software with a small recorded, random input (a seed value) from which any necessary random values can be computed in a deterministic, pseudorandom way. In this way, the decisionmaking process can be replayed so long as the seed is known and the randomness of the input is completely captured by the randomness of the seed. Using this technique, a decisionmaker would not have to generate a new random choice each time a random value is needed by a piece of software (such choices can be made by a cryptographic algorithm that uses the seed to yield reproducible values), nor know in advance how many random choices must be made. This technique allows software that makes random choices, such as a lottery, to be made fully reproducible and reviewable. Unlike capturing the entire environment, as was discussed above, this technique reduces the relevant portion of the environment to a very small and manageable value (the seed) while preserving the benefits of using randomness in the system.

If this technique is used, we also must prevent the decisionmaker from tampering with the seed value, as it fully determines all random data accessed by the program implementing the decision policy. Several methods can aid in ensuring the fair choice of seed values. A public procedure can be used to select a random value: for example, rolling dice or picking ping pong balls from the sort of device used by state lotteries.\textsuperscript{116} Alternatively, the seed value could be provided by a trusted third party, such as the random “beacon” operated by the U.S. National

\textsuperscript{115}In general, this approach will only find the top of some crest, which may or may not be the highest point on a hill (for instance, if a mountain has two peaks, one much higher than the other). Randomness helps fix this problem, however, since an algorithm can start climbing the hill at many different randomly chosen points and verify that they all reach the same highest point. Additionally, for many important problems, one can prove that only a certain limited number of optimal (i.e., highest or lowest) values exist. That is, if an analyst knows that the hill only has one peak, then which path a program takes to the top is irrelevant. For a description of the gradient descent approach to optimization and other approaches, see Richard O. Duda et al., Pattern Classification (2d ed. 2001).

\textsuperscript{116}Currently known strategies for generating public random values (“randomness beacons”) all have advantages as well as disadvantages--dice could be weighted; ping pong balls could be put in the freezer and the cold ones picked out of the machine. The National Institute of Standards and Technology runs a randomness beacon that has come under scrutiny because of distrust of the National Security Agency. To minimize these types of issues, the algorithm designer should pick the source of randomness most likely to be trusted by participants, which may vary. The algorithm designer could choose to collect many sources of random choices and mix them together to maximize the number of participants who will trust the randomness of the chosen seed. However, even physical sources of randomness that have not been tampered with have failed to be accountable for their goals in unexpected ways; for instance, the 1969 lottery for selecting draftees by birthday was later shown to be biased, with a disproportionate number of selectees coming from months early in the year. For a detailed overview of the problem and its causes, see Joan R. Rosenblatt & James J. Filliben, Randomization and the Draft Lottery, Science, Jan. 22, 1971, at 306.
Institute of Standards and Technology (NIST). In addition, it is possible for a set of mutually distrustful parties (possibly including decision subjects themselves) to engage in an interaction that produces a value that is unpredictable so long as at least one participant provided random input. Perhaps, the best option is to mix together randomness (sometimes called entropy) from many different sources. The simplest form of this practice would involve a decision subject entering a short random number as part of the input for their decision (e.g., on an application form). Then, the decisionmaker would generate a seed value for each decision by combining this known-to-the-subject, personal random value with (1) a pre-chosen random value to which the decisionmaker committed to using far in advance of seeing the personal random value, and (2) a unique identifier for the particular decision or decision subject that is difficult to change (e.g., the social security number of the participant). In order to foster maximum confidence that random choices are not improperly influenced, the decisionmaker should derive them using a combination of (1) a random value from a trusted third-party; (2) a random value chosen by the decisionmaker and possibly kept secret; (3) a participant- or decision-specific identifier that cannot be changed or controlled by the decisionmaker, such as a social security number, identification number, or other immutable piece of the subject’s name or data; and (4) a value chosen by the decisionmaker. Since these values are either outside of the decisionmaker’s control or are known, fixed, and subsequently verifiable before the inputs to a decision are known, using these

---

117 Computer science refers to a trusted third-party source of randomness as a “beacon.” The best known beacon is operated by the NIST, which publishes new random data every few minutes, ostensibly based on the measurement of quantum mechanical randomness via a device maintained in a NIST lab. NIST Randomness Beacon, Nat’l Inst. Standards & Tech. http://www.nist.gov/itl/csd/ct/nist_beacon.cfm [https://perma.cc/UNT3-6N6P] (last updated Sept. 21, 2016). Recent revelations about NIST’s role in allowing the U.S. National Security Agency to undermine the security of random number generation techniques standardized by NIST have led to some distrust of the NIST beacon, although it may be trustworthy in some applications. NIST standardized the Dual EC Deterministic Random Bit Generator (DUAL-EC) in SP 800-90A in 2007. At that time, cryptographers already knew the standard could accommodate a “backdoor,” or secret vulnerability. See Dan Shumow & Niels Ferguson, On the Possibility of a Back Door in the NIST SP800-90 Dual Ec Prng, in 7 Proc. Crypto (2007). Later, it was discovered that the NSA had very likely made use of this mechanism to create a backdoor in the standard itself. See Daniel J. Bernstein et al., Dual EC: A Standardized Back Door, in The New Codebreakers 256 (2016). Other beacon implementations have been proposed, including beacons based on “cryptocurrencies” such as Bitcoin. See, e.g., Joseph Bonneau et al., On Bitcoin as a Public Randomness Source, https://eprint.iacr.org/2015/1015.pdf [https://perma.cc/XQ38-FJ3H] (outlining a specific alternate proposal involving the use of Bitcoin as a source of publicly verifiable randomness).

118 Computer science has methods to simulate a trusted third party making a random choice. These methods require the cooperation of many mutually distrustful parties, such that as long as any one party chooses randomly, the overall choice is random. By selecting many participants in this process, one can maximize the number of people who will believe that the chosen value is in fact beyond undue influence. For an easy-to-follow introduction to these methods, see Manuel Blum, Coin Flipping by Telephone: A Protocol for Solving Impossible Problems, in 1981 Crypto, at 11
methods gives assurance that the decisionmaker is not skewing the results by controlling the selection of random values.\textsuperscript{119} Where random choices are part of a decisionmaking process, the fairness of the randomness (i.e., the consistency with the goal for which randomness was deployed in a particular system) used in those decisions should be verifiable. This can be achieved by relying on a small random seed and verifying its source. Once a random seed has been chosen in a satisfactory manner, it is still necessary to verify that the seed was in fact used in later decisions.\textsuperscript{120} This can be accomplished by the techniques we describe.

D. Applying Technical Tools Generally

Our general strategy in designing systems accountable for their procedural regularity is to require systems to create cryptographic commitments as digital evidence of their actions. Systems can be designed to publish commitments describing what they will do (i.e., a commitment to the decision policy enforced by the system, as represented by source code) before they are fielded and commitments describing what they actually did (i.e., a commitment to the inputs\textsuperscript{121} and outputs of any particular decision) after they are fielded. Zero-knowledge proofs can be used to ensure that these commitments actually correspond to the actions taken by a system.\textsuperscript{122} Indeed, it is possible to use zero-knowledge proofs to verify, for each decision, that the committed-to decision policy applied to the committed-to inputs yields the committed-to outputs.\textsuperscript{123} These zero-knowledge proofs could either be made public or provided to the system’s decision subjects along with their results.

\textsuperscript{119} When the fairness of random choices is key to the accountability of a decision process, great care must be taken in determining the source of random seed values, as many very subtle accountability problems are possible. For example, by changing the order in which decisions are taken, the decisionmaker can effectively “shop” for desirable random values by computing future deterministic pseudorandom values and picking the order of decisions based on its preference for which decisions receive which random choices. To prevent this, it may also be necessary to require that the decisionmaker take decisions in a particular order or that the decisionmaker commit to the order in which it will take decisions in advance of the seed being chosen. For a detailed description of the problems with randomness “shopping” and post-selection by a decision authority, see Joshua Alexander Kroll, Accountable Algorithms (Sept. 2015) (unpublished Ph.D dissertation, Princeton University) (on file with author).

\textsuperscript{120} For example, several state lotteries have been defrauded by insiders who were able to control what random values the lottery system used to decide winners. Specifically, an employee of the Multi-State Lottery Association (MUSL) was convicted of installing software on the system that controlled the random drawing and using the information gleaned by the software to purchase winning tickets for the association’s “Hot Lotto” game. See Grant Rodgers, Hot Lotto Rigger Sentenced to 10 Years, Des Moines Register (Sept. 9, 2015, 7:12 PM), http://www.desmoinesregister.com/story/news/crime-and-courts/2015/09/09/convicted-hot-lotto-rigger-sentenced-10-years/71924916 [https://perma.cc/U26A-8VMD] (describing the Iowa lottery fraud sentencing).

\textsuperscript{121} Note that, for these commitments to function, systems must also be designed to be fully reproducible, capturing all interactions with their environments as explicit inputs that can then be contained in published commitments. The use of seed values for randomization, discussed above in subsection II.C.4, offers one example of ensuring reproducibility.

\textsuperscript{122} The approach here was introduced in Kroll, supra note 119.

\textsuperscript{123} Id.
By disclosing commitments instead of source code or inputs and outputs, system operators can fully explain what their systems do without actually disclosing how those systems work up front. Later, if it becomes necessary to review the actions or decision policy of a system during a court case or regulatory action, system operators can disclose the contents of their commitments (that is, source code, inputs, and outputs), possibly under a protective regime. If it is possible to disclose these values publicly, then system operators may also choose (or be required) to do so. However, whether these data are disclosed or not, the published commitments and zero-knowledge proofs allow overseers and the public at large to verify that the decisions of some authority actually correspond to a specific predetermined policy rather than the arbitrary whim of a decisionmaker. Further, by observing that all decisions arise from the same policy, anyone reviewing these commitments can be certain that the policy was used for all decisions simply by checking that the commitments to the decision policy are consistent across decisions.

By requiring commitments to be published far in advance of any decision, it is possible to ensure that the particulars of a decision policy were chosen independently of the factors in the decisions it would render. For example, a decision policy that selects which individuals will receive a tax audit should be based on the risk of tax evasion, which in turn can be inferred by properties of the tax return itself. However, a corrupt tax authority could pick out individuals for audit and guess the particulars of their tax return data, then tailor the audit decision policy accordingly. Further, if a policy must be approved in advance by some oversight or certification body, the policy would need to be decided on and implemented in software far enough in advance to admit certification or review. Finally, if such certification does take place, subjects of the policy’s decisions (or overseers of those decisions) can be certain after the fact that the policy which was certified is the policy which was actually used in practice.¹²⁴

To the extent that the invariants of interest in a computer system are simple enough to compute, their truth can be verified by the same zero-knowledge proofs that attest to the relationship between the code, the inputs, and the outputs. Because powerful, modern zero-knowledge techniques can be applied to any code, they can also be applied to code that performs the analysis of these invariants, and the execution of that code can be considered as part of the operation of the system.¹²⁵

¹²⁴ Electronic voting systems have suffered from such problems in practice. In many jurisdictions, voting system software must be certified before it can be used in polling places. Systems are tested by the Election Assistance Commission (EAC), an independent commission created by the 2002 Help America Vote Act. See Testing and Certification Program, U.S. Election Assistance Commission, http://www.eac.gov/testing_and_certification [https://perma.cc/8DFX-LTYD] (detailing the EAC’s testing and certification regime). However, in many cases, updated, uncertified software has been used in place of certified versions because of pressure to include updated functionality or bug fixes. See, e.g., Fitrakis v. Husted, No. 2:12-cv-1015, 2012 WL 5411381 (S.D. Ohio Nov. 6, 2012) (involving a suit arising out of updates to voting systems immediately prior to the 2012 presidential election in Ohio).

¹²⁵ For example, suppose that we wish to demonstrate that a decision would be the same if the subject’s gender were reversed. The software implementing the decision could simply compute the
Simply publishing commitments to the inputs and outputs of a system rather than making them transparent will not solve all of the issues with transparency brought up in Part I. However, it can address the need for legitimate secrecy of the system, its inputs, or its outputs. Because it is possible using these methods to verify that a particular input and a particular decision policy correspond to a particular output, it is not strictly necessary to see these values in order to investigate the system’s procedural regularity.

We describe how certification of procedural regularity can be done for randomized software, such as software implementing a lottery, in greater detail in Section II.E below. Later, in Part III, we explain how these tools can extend to certify other, more complicated invariant properties of interest, enabling proof that a system comports with substantive goals or principles beyond simple procedural regularity.

E. Applying Technical Tools to Reform the Diversity Visa Lottery

Armed with these tools, we can turn to the question of how to ensure the procedural regularity of automated decisionmaking. To illustrate how designing a computer system can make it more accountable, we will apply the methods described above to a case study: the Diversity Visa Lottery (DVL) operated by the U.S. Department of State.

1. Current DVL Procedure

The DVL is run annually by the State Department to grant U.S. permanent resident visas (“green cards”) to 50,000 immigrants from around the world. The process, prescribed by 8 U.S.C. 1153(c), is intended to increase the national and regional diversity of immigrants to the U.S. by granting visas to a sample of people from countries otherwise underrepresented in the immigrant population.

The annual DVL process operates as follows. Would-be immigrants apply to be entered in the lottery, and applicants are grouped according to their country of birth. Within each country group, applicants are put into a rank-ordered list in a random order (the lottery step). The Attorney General then calculates the number of applicants to accept from each country using a formula based on the number of immigrants to the U.S. in recent years from each country and region. The calculated number of applicants is selected from the top of each country’s rank-ordered list. These “winners” are screened for eligibility to enter the U.S., and they receive visas if they are eligible. In some years, additional winners are selected so that all statutorily available visas can eventually be awarded, even if some applicants fail the screening process, drop out, or fail to proceed with their visa application.127

126 See generally Immigration and Nationality Act § 203(c), 8 U.S.C. § 1153(c) (2012); U.S. Dep’t of State, Foreign Affairs Manual ch. 9, § 502.6.
Questions have been raised about the correctness and accountability of this process. Would-be immigrants sometimes question whether the process is truly random or, as some suspect, is manipulated in favor of individuals or groups favored by the U.S. government. This suspicion, in turn, may subject DVL winners to reprisals, on the theory that winning the DVL is evidence of having collaborated secretly with U.S. agencies or interests.

There have also been undeniable failures in carrying out the DVL process. For example, the 2012 DVL initially reported incorrect results due to programming errors coupled with lax management.\textsuperscript{128} An accountable implementation of the DVL could address both issues by demonstrating that there is no favoritism in the process and by making it easy for outsiders to check that the process was executed correctly.

2. Transparency Is Not Enough

The DVL is an automated decision system for which transparency alone cannot solve its problems. First, the software implementing the decisions appears to be written in an irreproducible way.\textsuperscript{129} The system relies on the computer’s operating system to provide random numbers; thus, attempts to replicate the program’s execution at another time or on another computer will yield different random numbers and therefore a different DVL result. Notably, no amount of reading, analyzing, or testing of the software can remedy the nonreplicability of this software.

Second, the privacy interests of participants bar full transparency. People who apply to the DVL do not want their information, or even the fact that they applied, to be published. However, such publication is needed for the process to be verified through transparency and auditing. The Department could try to work around this problem by assigning an opaque record ID to each applicant and then having the lottery choose record IDs rather than applicants, but lottery operators could manipulate the outcome by retroactively assigning winning record IDs to people they wanted to favor. Further, it would be difficult to verify that no extra record IDs corresponding to actual participants had been added.

3. Designing the DVL for Accountability

Instead of this inherently unverifiable approach, we propose a technical solution for building an accountable version of the DVL.\textsuperscript{130} Using the techniques we have described, the State Department can demonstrate that it is running a fair lottery among a hidden set of participants.\textsuperscript{131}

\textsuperscript{129} Id.
\textsuperscript{130} A full technical analysis is beyond the scope of this paper.
\textsuperscript{131} Note that it is less straightforward to prove that the set of participants actually considered in the lottery matches the set of individuals who applied to be included. For example, the operator of the lottery might insert “shills,” or lottery entries that do not correspond to any real applicant, and if one of these applications were to be chosen, that place could be given to anyone of the Department’s choosing. It is technically nontrivial to prove that no extra applications were considered; studies of end-to-end secure voting protocols provide methods to do so. See, e.g., Daniel Sandler et al.,
We can solve the nonreplicability problem by choosing a random seed, as described previously. The third-party generating the random value used to create the seed could be one or more trusted NGOs, or applicants could provide a “PIN” on their applications.

Recall that the decision policy for the DVL is fixed in statute and hence publicly known. To provide additional oversight, the State Department could publish in the Federal Register a commitment to its software source code (far in advance of any decisions being made) and a commitment to all the inputs (i.e., to each data element in an application for the US visa) used to create the rank-ordered list and calculate the cut-off points. The State Department also could provide a zero-knowledge proof showing that applying the committed-to software to the committed-to inputs produces the announced lottery results. The proof could also demonstrate that the commitment to the software published in advance of all decisions is a commitment to the same software as the one used in each individual decision. These actions would bind the State Department to its choices of software, source code, and applicant data; ensure that the commitment to the software was not a fake; and prove that the same procedure was used to render each decision. Subsequent auditing by an oversight body could establish that the source code in the commitment faithfully implements the policy specified by statute (the code should be designed to enable this).

Finally, the State Department could determine an adequate method for generating a random seed to be used in the lottery step. This method should guarantee to the public that it is not possible for the State Department to choose winners by rearranging applications.\footnote{See supra note XX} This could be accomplished by combining random data chosen at a public ceremony (as is done for state lotteries); alternatively, the State Department could cooperate with interested NGOs to produce a verifiable random seed with a random value selected exclusively by the State Department (and published prior to the ceremony and any lottery applications) along with something that identifies a particular lottery entry uniquely (e.g., the applicant’s full application data, reduced by cryptographic hash, to a small numeric value). Depending on the implementation and application, the State Department could also include randomness selected by DVL applicants on their application, which could be harvested passively by tracking mouse movements during the application process.\footnote{Random choices in the DVL must be demonstrably random even to nonparticipants so that winners can plausibly claim that they were chosen by lottery and not because of sympathy for U.S. interests.}

Once these steps are taken, each applicant can be assured that the State Department’s decision on his application is fully explainable. If the applicant has questions regarding the process or a governmental overseer wants to audit it, the decisions will be replicable, and, if necessary, the secret source code and secret input data (including the random choices made in the lottery step) can be revealed

---

and verified--by a court or auditing agency--to be the proper code and data used to render the decision.\footnote{In fact, just as the applicant can be convinced that his decision is explainable without seeing the secret algorithm or secret inputs, an oversight body can be convinced that particular decisions were made correctly without seeing the applicant’s inputs, which might contain sensitive data, like health records or tax returns. Thus, subsequent auditing is rendered more useful and more acceptable to participants, as it can determine the basis for every decision without revealing sensitive information.}

These solutions depend on both redesigning the software code (a technical solution) and adopting procedures relating to how the software program is used (a legal or policy solution). They must be deployed during the design of the decision process and cannot salvage a poorly designed system after the fact.

In hindsight, it should not be surprising that the path to accountability for computational processes requires some redesign of the processes themselves. The same is true for noncomputational administrative processes, where the most accountable processes are those that are designed with accountability in mind.

III. Designing Algorithms to Assure Fidelity to Substantive Policy Choices

In Part I, we described methods that permit certification of properties of systems, and in Part II, we demonstrated how those methods can ensure that automated decisions are reached in accordance with agreed upon rules, a goal we called procedural regularity. In this Part, we examine how those methods could be used to certify other system properties that policymakers desire. Accountability demands not only that we certify that a policy was applied evenly across all subjects, but also that those subjects can be certain that the policy furthers other substantive goals or principles. A subject may want to know: Is the rule correctly implemented? Is it moral, legal, and ethical? Does it operate in the aggregate with fidelity to substantive policy choices?

We focus here on the goal of nondiscrimination\footnote{The word “discrimination” carries a very different meaning in engineering conversations than it does in public policy. Among computer scientists, the word is a value-neutral synonym for differentiation or classification: a computer scientist might ask, for example, how well a facial recognition algorithm successfully discriminates between human faces and inanimate objects. But, for policymakers, “discrimination” is most often a term of art for invidious, unacceptable distinctions among people—distinctions that either are, or reasonably might be, morally or legally prohibited. We use the latter meaning here.} in part because specific, additional technical tools have developed to assist with it and in part because the use of automated decisionmaking already has raised concerns about discrimination and the ability of current legal frameworks to deal with technological change.\footnote{See Pasquale, supra note 47, at 8-9 (describing the problem of discrimination through the use of automated decisionmaking).}

The well-established potential for unfairness in systems that use machine learning, in which the decision rule itself is not programmed by a human but rather inferred from data, has heightened these discrimination concerns. However, what makes a rule unacceptably discriminatory against some person or group is a fundamental and contested question. We do not address that question here, much less claim to
resolve it with computational precision. Instead, we describe how an emerging body of computer science techniques may be used to avoid outcomes that could be considered discriminatory.

Fidelity to policy choices like nondiscrimination is a more complicated goal than procedural regularity, and the solutions that currently exist to address it are less robust. Technical tools can ameliorate these problems, but they generally require a well-defined notion of what sort of fairness they are supposed to be enforcing. In this Part, we outline a few proposed well-defined notions. We present these techniques as examples of system properties that could be certified using the techniques described in Part I, but we do not necessarily advocate for any of them; ultimately, policymakers must decide whether these properties or others square with nondiscrimination goals.

In addition, the precision of computer code often brings into sharp focus the tensions surrounding antidiscrimination within current legal frameworks. Computers favor hard and fast rules over the types of standards and balancing tests often found in our common law system and civil rights law. While this tension suggests that doctrinal reform would make it easier to apply computerized decisionmaking in an area, we are not advocating a policy regime entirely made of bright line rules or predetermined fairness criteria. In fact, we believe that investigations of fairness should always be in the purview of ex post review processes. Instead, we offer an overview of the problem of algorithmic discrimination, the current state of the related technical tools, and the relationship of these tools to existing legal frameworks. Our aim is to both elucidate the current state of the art and suggest directions for further research and action.

A. Machine Learning, Policy Choices, and Discriminatory Effects

We focus here on decisions developed through machine learning—on situations where a machine has been “trained” through exposure to a large quantity of data and infers a rule from the patterns it observes. Computers are especially well-suited to discover patterns in these input-output pairs that can then guide future decisionmaking. In contrast to human-made rules, these rules for decisionmaking are induced from historical examples—they are, quite literally, rules learned by example. Humans orchestrate a computerized rule-creation process, rather than imparting the rules directly. These kinds of decisions raise problems for the methods described in Part I because they do not have obvious decision rules fixed by the system’s designer, which can be verified in the manner we have described. Instead, for the tools to apply, policymakers must determine that certain system properties would reflect substantive policy goals, and if such properties exist, they can then be certified and permit the type of accountability we have proposed above.

Machine learning is an increasingly common approach to solving problems that once seemed computationally intractable due to their complexity (e.g., object recognition in a photograph). The recent movement of software systems into a growing number of domains owes primarily to successful applications of machine learning, which is thus the primary focus of our analysis.

A significant concern about automated decisionmaking is that it may simultaneously systematize and conceal discrimination. Because it can be difficult
to predict the effects of a rule in advance (especially for large, complicated rules or rules that are machine-derived from data), regulators and observers may be unable to tell that a rule has discriminatory effects. In addition, decisions made by computers may enjoy an undeserved assumption of fairness or objectivity. However, the design and implementation of automated decision systems can be vulnerable to a variety of problems that can result in systematically faulty and biased determinations.

These decision rules are machine-made and follow mathematically from input data, but the lessons they embody may be biased or unfair nevertheless. Below, we describe a few illustrative ways that models, or decision rules derived from data, generated through machine learning, may turn out to be discriminatory. We adapt a taxonomy laid out in previous work by Solon Barocas and Andrew D. Selbst and make use of the “catalog of discriminatory evils” of machine learning systems laid out by Hardt and Dwork et al.

First, algorithms that include some type of machine learning can lead to discriminatory results if the algorithms are trained on historical examples that reflect past prejudice or implicit bias, or on data that offer a statistically distorted picture of groups comprising the overall population. Tainted training data would be a problem, for example, if a program to select among job applicants is trained on the previous hiring decisions made by humans, and those previous decisions were themselves biased. Statistical distortion, even if free of malice, can produce similarly troubling effects: consider, for example, an algorithm that instructs police to stop and frisk pedestrians. If this algorithm has been trained on a dataset that overrepresents the incidence of crime among certain groups (because these groups have historically been the target of disproportionate enforcement), the algorithm may direct police to detain members of these groups at a disproportionately high rate (and nonmembers at a disproportionately low rate). Such was the case with the

---

137 See Paul Schwartz, Data Processing and Government Administration: The Failure of the American Legal Response to the Computer, 43 Hastings L.J. 1321, 1342 (1992) (describing the deference that individuals give to computer results as the “seductive precision of output”).

138 See id. at 1342-43 (noting that the computer creates “new ways to conceal ignorance and subjectivity” because people overestimate its “accuracy and applicability”).

139 See Barocas & Selbst, supra note 8 (describing a taxonomy that isolates specific technical issues to create a decisionmaking model that may disparately impact protected classes).


142 See Barocas & Selbst, supra note 8, at 682 (citing Stella Lowry & Gordon Macpherson, A Blot on the Profession, 296 Brit. Med. J. 657, 657 (1988)) (describing how a hospital developed a computer program to sort medical school students based on previous decisions that had disfavored racial minorities and women). Another example is a Google algorithm that showed ads for arrest records much more frequently when black-identifying names were searched than when white-identifying names were searched—likely because users clicked more often on arrest record ads for black-identifying names and the algorithm learns from this behavior with the purpose of maximizing click-throughs. Id. at 682-83 (citing Latanya Sweeney, Discrimination in Online Ad Delivery, Comm. ACM, May 2013, at 44, 47 (2013)).
New York City Police Department’s stop-and-frisk program, for which data from 2004 to 2012 showed that 83% of the stops were of black or Hispanic persons and 10% were of white persons in a resident population that was 52% black or Hispanic and 33% white.\footnote{David Rudovsky & Lawrence Rosenthal, \textit{Debate: The Constitutionality of Stop-and-Frisk in New York City}, 162 U. Pa. L. Rev. Online 117, 120-21 (2013).} Note that the overrepresentation of black and Hispanic people in this sample may lead an algorithm to associate typically black or Hispanic traits with stops that lead to crime prevention, simply because those characteristics are overrepresented in the population that was stopped.\footnote{The underrepresentation of white people would likely cause the opposite effect, though it could be counter-balanced if, say, the police stopped a subset of white people who were significantly more likely to be engaged in criminal behavior.}

Second, machine learning models can build in discrimination through choices in how models are constructed. Of particular concern are choices about which data models should consider, a problem computer scientists call \textit{feature selection}. Three types of choices about inputs could be of concern: using membership in a protected class directly as an input (e.g., decisions that take gender into account explicitly); considering an insufficiently rich set of factors to assess members of a protected class with the same degree of accuracy as nonmembers (e.g., in a hiring application, if fewer women have been hired previously, data about female employees might be less reliable than data about male employees); and relying on factors that happen to serve as proxies for class membership (e.g., women who leave a job to have children lower the average job tenure for all women, causing this metric to be a known proxy for gender in hiring applications). Eliminating proxies can be difficult, because proxy variables often contain other useful information that an analyst wishes the model to consider (for example, zip codes may indicate both race and differentials in local policy that is of legitimate interest to a lender). The case against using a proxy is clearer when alternative inputs could yield equally effective results with fewer disadvantages to protected class members. A problem of insufficiently rich data might be remedied in some cases by gathering more data or more features, but if discrimination is already systemic, new data will retain the discriminatory impact. While it is tempting to say that technical tools could allow perfect enforcement of a rule barring the use of protected attributes, this may in fact be an undesirable policy regime. As previously noted, there may be cases where allowing an algorithm to consider protected class status can actually make outcomes fairer. This may require a doctrinal shift, as, in many cases, consideration of protected status in a decision is presumptively a legal harm.

Third and finally, there is the problem of “masking”: intentional discrimination disguised as one of the above-mentioned forms of unintentional discrimination. A prejudiced decisionmaker could skew the training data or pick proxies for protected classes with the intent of generating discriminatory results.\footnote{See Barocas & Selbst, supra note 8, at 692-93 (describing ways to intentionally bias data collection in order to generate a preferred result).}
More pernicious masking could occur at the level of designing a machine learning model, which is a very human-driven, exploratory process.\footnote{In other words, the machine learning model would be intentionally coded to develop bias.}

\section*{B. Technical Tools for NonDiscrimination}

As mentioned in the previous Part, transparency and after-the-fact auditing can only go so far in preventing undesired results. Ideally, those types of ex post analyses should be used in tandem with powerful ex ante techniques during the design of the algorithm. The general strategy we proposed in Section II.D--publishing commitments and using zero-knowledge proofs to ensure that commitments correspond to the system’s decisionmaking actions--can certify any property of the decision algorithm that can be checked by a second examination algorithm.\footnote{Such an algorithm might be a tool for verifying properties of software or simply a software test. See \textsuperscript{supra} Part I (discussing software testing and software verification in greater detail).} Such properties can be proven by making the examination algorithm public and giving a zero-knowledge proof that, if the examination algorithm were run on the secret decision algorithm, it would report that the decision algorithm has the desired property. The question then is which, if any, properties policymakers would want to build into particular decision systems.

A simple example of such a property would be the exclusion of a certain input from the decisionmaking process. A decisionmaker could show that a particular algorithm does not directly use sensitive or prohibited classes of information, such as gender, race, religion, or medical status.

The use of machine learning adds another wrinkle because decision rules evolve on the fly. However, the absence of static, predetermined decision rules does not necessarily preclude the use of our certification strategy. Computer scientists, including Hardt,\footnote{Hardt, \textsuperscript{supra} note 140.} Dwork et al.,\footnote{Dwork et al., \textsuperscript{supra} note 141.} and others, have developed techniques that formalize fairness in such a way that they can constrain the machine learning process so that learned decision rules have specific well-defined fairness properties. These properties also can be incorporated in the design of systems such that their inclusion in the decisionmaking process can be certified and proven.\footnote{We concentrate on certification and proof of a system property to an overseer, observer, or participant. However, these tools are also valuable for compliance (since proofs can certify to the implemen ter of a system that the system is working as intended) and for demonstration that a decisionmaker will be able to show how and why they used certain data after the fact in case of an audit or dispute.}

We describe three such properties below. First, decisions can incorporate randomness to maximize the gain of learning from experience. Second, computer science offers many emerging approaches to maximize fairness, defined in a variety of ways, in machine learning systems. At a high level, all of these definitions reduce to the proposition that similarly situated people should be treated similarly, without regard to sensitive attributes. As we shall see, simple blindness to these attributes is not sufficient to guarantee even this simplified notion of fairness. Finally, related ideas from differential privacy can also be used to guarantee that protected status could not have been a substantial factor in certain decisions.
Those who use algorithmic decisionmaking today regularly make assertions about properties of these algorithms without proving them. This likely occurs because they are required by law to disclose certain facts about their decision process to regulators and consumers,\(^{151}\) they simply want to generate good will,\(^{152}\) or they demonstrate better behavior than a competitor.\(^{153}\) But without proof, these assertions are just words on paper, subject to challenge by skeptical regulators and disbelief by skeptical consumers. This skepticism is not entirely unfounded: these assertions have proved false in the past.\(^{154}\) Digital evidence such as zero-knowledge proofs gives a direct connection between the fact being asserted and the technical mechanism of decisionmaking. This proof provides the consumer with a high assurance that the assertion proffered relates meaningfully to the facts on the ground.

1. Learning from Experience

As mentioned in Section I.B, incorporating randomness into an algorithm can give it flexibility to operate outside of the environment for which it was designed. Similarly, randomness can prevent hidden biases in the design or deployment of an algorithm from leading to consistent discriminatory outcomes. There is a large and rich literature on how to maximally learn from previous data and how to use random choices to ensure that a model is as faithful as possible to the real world.\(^{155}\)

Consider a machine learning algorithm for hiring that is trained using a biased set of initial data indicating that women are weak candidates, even though gender does not predict job performance among the full population. If the resulting model would hire mostly men, the algorithm for hiring can create a self-fulfilling prophecy in which it finds that characteristics of successful hires correlate strongly with proxies for gender. But, if the algorithm is designed to incorporate an element of randomness such that some candidates who are not predicted to do well get hired (and have their performance tracked), the validity of the initial assumptions can be tested and the accuracy and fairness of the entire system will benefit over time. By occasionally guessing about candidates for which the model cannot make confident predictions, the model can gather additional data and evolve to become more faithful to the real world.

Similarly, randomness is often necessary when training machine learning models. Models may become too specialized or specific to the data used for training, a problem called “overfitting.” Randomness can prevent this problem. Likewise, models may find a decision rule is well-suited for some portion of the input, but not the best rule overall. Randomness can also help avoid this bias. Consider, for example, a credit-scoring model trained initially on a biased set of

\(^{151}\) See, e.g., 12 C.F.R §§ 203.4-5 (2015) (providing requirements for the compilation, disclosure, and reporting of loan data).

\(^{152}\) [CITE example TK]

\(^{153}\) [CITE example TK]

\(^{154}\) [CITE example TK]

\(^{155}\) This literature is divided between the machine learning research community in Computer Science and the study of optimal decisionmaking in Statistics. See supra note 17.
data that underrates the creditworthiness of some minority group. Even if the model is the best possible decision rule for a population matching the biased input data, the model may unfairly deny access to credit to members of that biased minority group. In addition to the discrimination, the use of this model would deny creditors business opportunities with the unfairly rejected individuals. Here again, allowing the model to occasionally guess randomly, while tracking expected versus actual performance, can improve the model’s faithfulness to the population on which it is actually used rather than the biased population on which it was trained. The information from this injection of randomness can be fed back to the model to improve the accuracy and fairness of the system overall.

2. Fair Machine Learning

One commonly understood way to demonstrate that a decision process is independent of sensitive attributes is to preclude the use of those sensitive attributes from consideration. For example, race, gender, and income may be excluded from a decisionmaking process to assert that the process is “race-blind,” “gender-blind,” or “income-blind.”156 From a technical perspective, however, this approach is naive. Blindness to a sensitive attribute has long been recognized as an insufficient approach to making a process fair. The excluded or “protected” attributes can often be implicit in other nonexcluded attributes. For example, when race is excluded as a valid criterion for a credit decision, redlining may occur when a zip code is used as proxy that closely aligns with race.157

This type of input “blindness” is insufficient to assure fairness and compliance with substantive policy choices. Although there are many conceptions of what fairness means, we consider here a definition of fairness in which similarly situated people are given similar treatment—that is, a fair process will give similar participants a similar probability of receiving each possible outcome. This is the core principle of a developing literature on fair classification in machine learning, an area first formalized by Dwork, Hardt, Pitassi, Reingold, and Zemel.158 This work stems from a longer line of research on mechanisms for data privacy.159 We further describe the relationship between fairness in the use of data and privacy below.

156 See, e.g., 12 C.F.R. § 1002.5(b) (2015) (“A creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction.”); 12 C.F.R. § 1002.6(b)(9) (2015) (“[A] creditor shall not consider race, color, religion, national origin, or sex (or an applicant’s or other person’s decision not to provide the information) in any aspect of a credit transaction.”)
158 Dwork et al., supra note 141.
159 Specifically, the work of Dwork et al. is a generalization of ideas originally presented in Cynthia Dwork, Differential Privacy, 2006 Proc. 33rd Int’l Colloquium on Automata, Languages & Programming 1. As discussed below, fairness can be viewed as the property that sensitive or protected status attributes cannot be inferred from decision outcomes, which is very much a privacy property.
The principle that similar people should be treated similarly is often called individual fairness and it is distinct from group fairness in the sense that a process can be fair for individuals without being fair for groups.\(^{160}\) Although it is almost certainly more policy-salient, group fairness is more difficult to define and achieve. The most commonly studied notion of group fairness is statistical parity, the idea that an equal fraction of each group should receive each possible outcome. While statistical parity seems like a desirable policy because it eliminates redundant or proxy encodings of sensitive attributes, it is an imperfect notion of fairness. For example, statistical parity says nothing about whether a process addresses the “right” subset of a group. Imagine an advertisement for an expensive resort: we would not expect that showing the advertisement to the same number of people in each income bracket would lead to the same number of people clicking on the ad or buying the associated product. For example, a malicious advertiser wishing to exclude a minority group from a resort could design its advertising program to maximize the likelihood of conversion for the desired group while minimizing the likelihood that the ad will result in a sale to the disfavored group. In the same vein, if a company aimed to improve the diversity of its staff by offering the same proportion of interviews to candidates with minority backgrounds as are minority candidate applications, that is no guarantee that the number of people hired will reflect the population of applicants or the population in general. And the company could hide discriminatory practices by inviting only unqualified members of the minority group to apply, effectively creating a self-fulfilling prophecy for decision rules established by machine learning.

The work of Dwork et al. identifies an additional interesting problem with the “fairness through blindness” approach: by remaining blind to sensitive attributes, a classification rule can select exactly the opposite of what is intended.\(^{161}\) Consider, for example, a system that classifies profiles in a social network as representing either real or fake people based on the uniqueness of their names. In European cultures, from which a majority of the profiles come, names are built by making choices from a relatively small set of possible first and last names, so a name which is unique across this population might be suspected to be fake. However, other cultures (especially Native American cultures) value unique names, so it is common for people in these cultures to have names that are not shared with anyone else. Since a majority of accounts will come from the majority of the population, for which unique names are rare, any classification based on the uniqueness of names will inherently classify real minority profiles as fake at a higher rate than majority profiles,\(^{162}\) and may also misidentify fake profiles using names drawn from the minority population as real. This unfairness could be remedied if the system were “aware” of the minority status of a name under

\(^{160}\) Sometimes, a more restrictive notion of individual fairness implies group fairness. Id. Intuitively, this is because if people who are sufficiently similar are treated sufficiently similarly, there is no way to construct a minority of people who are treated in a systematically different way.

\(^{161}\) See Dwork et al., supra note 141.

\(^{162}\) That is, the minority group will have a higher false positive rate.
consideration, since then the algorithm could know whether the implication of a unique name is that a profile is very likely to be fake or very likely to be real.\footnote{In this case, differential treatment based on a protected status attribute improves the performance of the automated decision system in a way that requires that the system know and make use of the value of that attribute.}

This insight explains why the approach taken by Dwork et al. is to enforce similar probabilities of each possible outcome on similar people, requiring that the aggregate difference in probability of any individual receiving any particular outcome be limited.\footnote{See Dwork et al., supra note 141, at 215 (explaining that fairness can be captured under the principle that “two individuals who are similar with respect to a particular task should be classified similarly”).} Specifically, Dwork et al. require that this difference in chance of outcome be less than the difference between individuals subject to classification.\footnote{This is formalized as the proposition that the difference in probability distributions between outcomes for each subgroup of the population being classified is less than the difference between those groups, for a suitable measurement of the difference between groups. For technical reasons, this particular formulation is mathematically convenient, although different bounds might also be useful. For the formal mathematical definition, see Dwork et al., supra note 141, at 216.} This requires a mathematically precise notion of how “different” people are, which might be a score of some kind or might naturally arise from the data in question.\footnote{For example, if the physical location of subjects is a factor in classification, we might naturally use the distance between subjects as one measure of their similarity.} This notion of similarity must also capture all relevant features, including possibly sensitive or protective attributes such as minority status, gender, or medical history. Because this approach requires the collection and explicit use of sensitive attributes, the work describes its definition of fairness as fairness through awareness.\footnote{\textit{Id.} at 215.} While the work of Dwork et al. provides only a theoretical framework for building fair classifiers, others have used it to build practical systems that perform almost as well as classifiers that are not modified for fairness.

The work of Dwork et al. also provides the theoretical basis for a notion of \textit{fair affirmative action}, the idea that imposing an external constraint on the number of people from particular subgroups who are given particular classifications should have a minimal impact on the principle that similar people are treated similarly. This provides a technique for forcing a fairness requirement such as statistical parity even when it will not arise naturally from some classifier.

A more direct approach to making a machine learning process fair is to modify or select the input data in such a way that the output satisfies some fairness property. For example, in order to make sure that a classifier does not over-reflect the minority status of some group, we could select extra training samples from that group or duplicate samples we already have. In either case, care must be taken to avoid biasing the training process in some other way or overfitting the model to the nonrepresentative data.

Other work focuses on \textit{fair representations} of data sets. For example, we can take data points and assign them to \textit{clusters}, or groups of close-together points, treating each cluster as a prototypical example of some portion of the original data
set. This is the approach taken by Zemel, Wu, Swersky, Pitassi, and Dwork.\textsuperscript{168} Specifically, Zemel et al. show how to generate such prototypical representations automatically and in a way that guarantees statistical parity for any subgroup in the original data. In particular, the probability that any person in the protected group is mapped to any particular prototype is equal to the probability that any person not from the protected group is mapped to the same prototype.\textsuperscript{169} Therefore, classification procedures which have access only to the prototypes must necessarily not discriminate, since they cannot tell whether the prototype primarily represents protected or unprotected individuals. Zemel et al. test their model on many realistic data sets, including the Heritage Health Prize data set, and determine that it performs nearly as well as best-of-breed competing methods while ensuring substantial levels of fairness.\textsuperscript{170} This technique allows for a kind of “fair data disclosure,” in which disclosing only the prototypes allows any sort of analysis, fair or unfair, to be run on the data set to generate fair results.

A related approach is to use a technique from machine learning called regularization, which involves introducing new information to make trained models more generalizable in the form of a penalty assigned to undesirable model attributes or behaviors. This approach has also led to many useful modifications to standard tools in the machine learning repertoire, yielding effective and efficient fair classifiers.\textsuperscript{171}

The work of Zemel et al. suggests a related approach, which is also used in practice: the approach of generating fair synthetic data. Given any set of data, we can generate new data such that no classifier can tell whether a randomly chosen input was drawn from the real data or the fake data. Furthermore, we can use approaches like that of Zemel et al. to ensure that the new data are at once representative of the original data and also fair for individuals or subgroups. Because synthetic data are randomly generated, they are useful in situations where training a classifier on real data would create privacy concerns. Also, synthetic data can be made public for others to use, although care must be taken to avoid allowing others to infer facts about the underlying real data. Such model inversion attacks\textsuperscript{172} have been demonstrated in practice, along with other inference or deanonymization attacks that allow sophisticated conclusions without direct access to the actual data that give rise to the conclusions.\textsuperscript{173}

\textsuperscript{168} Richard Zemel et al., Learning Fair Representations, 28 Proc. 30th Int’l Conf. on Machine Learning 325 (2013).

\textsuperscript{169} Id.

\textsuperscript{170} Id.

\textsuperscript{171} See, e.g., Toshihiro Kamishima et al., Fairness-Aware Learning Through Regularization Approach, 2011 Proc. 3rd IEEE Int’l Workshop on Privacy Aspects of Data Mining 643 (describing a model in which two types of regularizers were adopted to enforce fair classification).

\textsuperscript{172} See Matthew Fredrikson et al., Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, 2014 Proc. 23rd USENIX Security Symp. 17 (describing privacy risks in which attackers can predict a patient’s genetic markers if provided with the model and some demographic information).

\textsuperscript{173} For an overview of these techniques, see Arvind Narayanan & Edward W. Felten, No Silver Bullet: De-identification Still Doesn’t Work (July 9, 2014),
All of these approaches demonstrate that it is possible to build a wide variety of definitions of fairness into a wide variety of data analysis and classification systems, at least to the extent that a definition of fairness is known or can be approximated in advance. There are no bright-line rules that allow the designer or operator of a machine learning system to guarantee that the system’s behavior is compliant with antidiscrimination laws. Nor do we believe that such rules can or even should exist. It is not for technologists to define an ex ante notion of fairness that applies in all cases or even just for a specific system. Rather, fairness must be determined contextually and often must be reviewed ex post. Regardless, it is certainly not impossible to build fairness into automated decision systems, which shows that unconstrained use of data analysis is not always necessary. Uses of data that do not employ methods to investigate or ensure fairness must account for their decision policies in some other way.

Many of these approaches rely on the insufficient notion of group fairness by statistical parity. To the extent that more technical research can help to address the problem of unfairness in big data analysis, it is by expanding the repertoire of definitions of group fairness that can be usefully applied in practice and by providing better exploratory and explanatory tools for comparing different notions of fairness. From a public policy perspective, it would be extremely useful to system designers to have a set of rules, standards, or best practices that explain what notions of fairness should be used in specific real-world applications.

A complementary notion to machine learning systems that can guarantee prespecified, formal fairness properties is the work of Rudin on machine learning systems that are interpretable.\footnote{Cynthia Rudin, Algorithms for Interpretable Machine Learning, 2014 20th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining 1519.} Such systems generate models that can be used to classify individuals, but also explanations for why those classifications were made. These explanations can be reviewed later to understand why the model behaves a certain way, and in some cases how changes in the input data would affect the model’s decision. These explanations can be extremely valuable to experts and oversight authorities, who wish to avoid treating models as black boxes.

3. Discrimination, Data Use, and Privacy

A different way to define whether a classification is fair is to say that we cannot tell from the outcome whether the subject was a member of a protected group or not. That is, if an individual’s outcome does not allow us to predict that individual’s attributes any better than we could by guessing them with no information, we can say that outcome was assigned fairly. To see why this is so, observe the contrary: if the fact that an individual was denied a loan from a particular bank tells you that this individual is more likely to live in a certain neighborhood, this implies that you hold a strong belief that the bank denies credit to residents of this neighborhood and hence a strong belief that the bank makes

http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

174 Cynthia Rudin, Algorithms for Interpretable Machine Learning, 2014 20th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining 1519.
decisions based on factors other than the objective credit risk presented by applicants.

Thus, fairness can be seen as a form of an information hiding requirement similar to privacy. If we accept that a fair decision does not allow us to infer the attributes of a decision subject, we are forced to conclude that fairness is protecting the privacy of those attributes.

Indeed, it is often the case that people are more concerned that their information is used to make some decision or classify them in some way than they are that the information is known or shared. This concern relates to the famous conception of privacy as the “right to be left alone,” in that generally people are concerned with the idea that disclosure interrupts their enjoyment of an “inviolable personality.”

Data use concerns also surface in the seminal work of Solove, who refers to concerns about “exclusion” in “information processing,” or the lack of disclosure to and control by the subject of data processing and “distortion” of a subject’s reputation by way of “information dissemination.” Solove argues that these problems can be countered by giving subjects knowledge of and control over their own data. In this framework, the predictive models of automated systems, which might use seemingly innocuous or natural behaviors as inputs, create anxiety on the part of data subjects. We propose a complementary approach: if a system’s designer can prove to an oversight entity or to each data subject that the sorts of behaviors that cause these anxieties are simply not possible behaviors of the system, the use of these data will be more acceptable.

We can draw an analogy between data analysis and classification problems and the more familiar data aggregation and querying problems that are much discussed in the privacy literature. Decisions about an individual represent (potentially private) information about that individual (i.e., one might infer the input data from the decision), and this raises concerns for privacy. In essence, privacy may be at risk from an automated decision that reveals sensitive information just like fairness may be at risk from an automated decision. In this analogy, a vendor or agency using a model to draw automated decisions wants those decisions to be as accurate as possible, corresponding to the idea in privacy that it is the goal of a data analyst to build as complete and accurate a picture of the data subject as is feasible.

A naive approach to making a data set private is to delete “personally identifying information” from the data set. This is analogous to the current practice of making data analysis fair by removing protected attributes from the input data. However, both approaches fail to provide their promised protections. The failure

---

177 Id. at 546 (detailing privacy statutes that allow individuals to access and correct information that is maintained by government agencies).
178 Reidentification of individuals based on inferences from disparate data sets is a growing and important concern that has spawned a large literature in both Computer Science and Law. See Ohm, supra note 72, 1704 (arguing that developments in computer science demonstrate that “[d]ata can
in fairness is perhaps less surprising than it is in privacy—discrimination law has known for decades about the problem of proxy encodings of protected attributes and their use for making inferences about protected status that may lead to adverse, discriminatory effects.\footnote{For example, the law explicitly forbids the (sole) use of certain attributes that are likely to be highly correlated with protected status categories, as in protections against redlining. See, e.g., 12 C.F.R. § 1002.5(b) (2015) (“A creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction”); 12 C.F.R. § 1002.6(b)(9) (2015) (“[A] creditor shall not consider race, color, religion, national origin, or sex (or an applicant’s or other person’s decision not to provide the information) in any aspect of a credit transaction.”).}

The work of Hardt\footnote{Hardt, \textsuperscript{\textsuperscript{180}}} relates the work on fairness by Dwork et al.\footnote{Dwork et al., \textsuperscript{\textsuperscript{141}}} to the work on differential privacy by Dwork.\footnote{Dwork, \textsuperscript{\textsuperscript{159}}} As differential privacy is a well-founded notion of protection against inferences and the recovery of an individual identity from “anonymous” data, so are formal fairness properties for automated decision systems sound notions of fairness for individuals and a theoretical framework on which to ground more complicated notions of fairness for protected groups.

The many techniques of building fair data analysis and classification systems described above mostly require decisionmakers to have access to protected status information, at least during the design phase of the algorithm. However, in many cases, concerns about misuse, reuse, or abuse of this information have led to a policy regime where decisionmakers are explicitly barred from using such information. The deployment of these technical tools would require a policy change.\footnote{One example is the privacy regime created by the Health Insurance Portability and Accountability Act, see \textsuperscript{\textsuperscript{79}} supra note 79, which forbids the disclosure of certain types of covered information beyond those for which the data subject was previously given notice and which limits disclosure to covered entities subject to the same restrictions.} The techniques described above could be used to make such a change less prone to engendering the very real concerns of data abuses that have led to the current regime.

C. AntiDiscrimination Law and Algorithmic Decisionmaking

The goal of Part II—procedural regularity—is relatively simple from a legal standpoint. Procedural regularity is a core idea behind due process: the state cannot single out an individual for a different procedure.\footnote{See, e.g., Arthur S. Miller, \textit{An Affirmative Thrust to Due Process of Law?}, 30 Geo. Wash. L. Rev. 399, 403 (1962) (“Procedural due process (‘adherence to procedural regularity’), as we have often been told by Supreme Court justices, is the very cornerstone of individual liberties.”).} An argument that governance measures ensuring algorithmic procedural regularity are required by due process is more tenuous,\footnote{See Citron, \textsuperscript{\textsuperscript{\textsuperscript{159}}} supra note 6, at 1278-1300 (arguing that current procedural protections are inadequate for automated decisionmaking).} but an agency that implements such measures will not risk violating a legal requirement.

\footnote{\textit{be either useful or perfectly anonymous but never both}, and that such developments should “trigger a sea change” in legal scholarship).}
In contrast, governance of algorithms to promote nondiscrimination runs into the complicated field of antidiscrimination law. Here, the movement toward a colorblind interpretation of equal protection has created friction with the precedents involving disparate impact. We argue that, given the current state of antidiscrimination law, designing for nondiscrimination is important because users of algorithms may be legally barred from revising processes to correct for discrimination after the fact, and technical tools offer solutions to help.


Antidiscrimination law is based upon both the constitutional guarantee of equal protection and supplemental statutory protections. Modern interpretations of the Equal Protection Clause generally have been divided into two camps: those who believe in a color-blind Constitution—protecting individualized assessments and eschewing any evaluations based on group status—and those who support antisubordination attempts to remedy inequalities between groups. The general trend has been toward colorblindness.

For statutory measures, we will focus on Title VII of the Civil Rights Act of 1964. Under Title VII, remedies are available for disparate treatment--discriminatory intent or the formal application of different rules to people of different groups--and disparate impact--results that differ for different groups.

---

186 See U.S. Const. amend. XIV, § 1 (“No State shall . . . deny to any person within its jurisdiction the equal protection of the laws.”). The Equal Protection Clause has also been interpreted to apply to the federal government through the Due Process Clause of the Fifth Amendment. See, e.g., Kenji Yoshino, The New Equal Protection, 124 Harv. L. Rev. 747, 748 n.10 (2010).

187 See, e.g., Reva B. Siegel, From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases, 120 Yale L.J. 1278, 1281 (2011) (describing this binary as the common interpretation of equal protection jurisprudence).

188 See The Supreme Court, 2008 Term--Leading Cases, 123 Harv. L. Rev. 153, 289 (2009) (“The Court’s conception of equal protection turns largely on its swing voter, Justice Kennedy, who appears to support a moderate version of the colorblind Constitution.”). But see Reva B. Siegel, The Supreme Court, 2012 Term--Foreword: Equality Divided, 127 Harv. L. Rev. 1, 6 (2013) (agreeing that “[s]hifts in equal protection oversight . . . are continuing to grow” but arguing that these changes are “neither colorblind nor evenhanded” because “the Court has encouraged majority claimants to make discriminatory purpose arguments about civil rights law based on inferences the Roberts Court would flatly deny if minority claimants were bringing discriminatory purpose challenges to the criminal law”).


Algorithmic decisionmaking blurs the definitions of disparate treatment and disparate impact and poses a number of open questions.191

Is it disparate treatment when the inputs used are a proxy for membership in a protected class? Different rules are effectively applied to different groups in this case, but that difference may have no effect on the outcomes.192 If the people responsible for a decision know that an algorithm behaves in a way that has disparate impact, does that mean that they intend a discriminatory result?193 If an algorithm generates poor outcomes for a group of people, how accurate does the algorithm need to be (and how carefully does the decisionmaker need to test alternative algorithms) before the decisionmaker can escape disparate impact liability because the factors used are job-related?194 If, as noted in subsection III.B.2, knowledge of class membership can be used to improve the fairness of outcomes for members of all classes, should doing so be considered disparate treatment?

These doctrines were recently considered in Ricci v. DeStefano, in which the Supreme Court held that “before an employer can engage in intentional discrimination for the asserted purpose of avoiding or remedying an unintentional disparate impact, the employer must have a strong basis in evidence to believe it will be subject to disparate-impact liability if it fails to take the race-conscious, discriminatory action.”195 At issue was the City of New Haven’s test for firefighter promotions; though the tests had been constructed in an attempt to ensure there was no discrimination by race,196 the pass rates for minorities were about half of the pass rate for whites.197 The New Haven Civil Service Board did not certify the results of the test (and validate the promotions) due to concerns about fairness and disparate impact liability for the City.198

Ricci demonstrates the tension between disparate treatment and disparate impact. Facialy neutral policies can produce unequal results for protected classes, but remediing that disparate impact would require the state to treat people differently based on class membership, which Ricci forbids. Ricci also hints at the difficulties in squaring the Court’s move towards a colorblind interpretation of the Equal Protection Clause and the doctrine of disparate impact. The holding does not directly address the constitutional issue, but Justice Scalia’s concurrence does note that the “war between disparate impact and equal protection will be waged sooner or later.”199 Both of these doctrinal tensions are of concern to lawmakers and policymakers.

---

191 See Barocas & Selbst, supra note 8, at 694–714 (noting the ways in which algorithmic data mining techniques can lead to unintentional discrimination against historically prejudiced groups).
192 Id. at 695.
193 Id. at 700.
194 Id. at 707.
196 Id. at 565.
197 Id. at 586-87.
198 Id. at 579.
199 Id. at 595-96 (Scalia, J., concurring).
2. **Ricci Impels Designing for NonDiscrimination**

Although Ricci has generated wide-ranging conversation about equal protection, disparate treatment, and disparate impact, we wish to emphasize its implications for the governance of decision algorithms for processes where nondiscrimination is a goal. The holding in Ricci suggests that we cannot solely rely on auditing for legal reasons in addition to the reasons discussed in Section II.A. If an agency runs an algorithm that has a disparate impact, correcting those results after the fact will trigger the same kind of analysis as New Haven’s rejection of its firefighter test results. It is even possible that the Court will “subject some range of disparate impact compliance efforts to strict scrutiny,”200 a high bar that will be difficult to satisfy in most cases.

The legal difficulties with correcting discriminatory algorithms ex post make measures to design algorithms for nondiscrimination even more important. The Court in Ricci took no issue with New Haven’s process of designing the tests with an eye towards nondiscrimination, reasoning that “Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race.”201 However, “once that process has been established and employers have made clear their selection criteria, they may not then invalidate the test results, thus upsetting an employee’s legitimate expectation not to be judged on the basis of race.”202

The uneasy fit of algorithmic decisionmaking into the disparate treatment/disparate impact framework does mean that someone could allege disparate treatment because the design of the algorithm includes inputs that are a proxy for class membership, resulting in a formal application of different rules to different groups of people. However, such a claim would be valid against virtually any system with a significant number of inputs. It seems more likely that courts would reject the formal-rule subset of disparate treatment for algorithmic decisions than that they would hold the majority of algorithmic decisionmaking to be disparate treatment. In the end, incorporating nondiscrimination in the initial design of algorithms is the safest path that decisionmakers can take, and we should encourage the development and deployment of technical tools to aid in that design.

IV. **Fostering Collaboration Across Computer Science, Law, and Policy**

In this Part, we consider how the types of technological assurance described in previous Parts relate to mechanisms of oversight in law and public policy. In technical approaches, it is traditional to have a detailed, well-defined specification of the behavior of a system for all types of situations. In lawmaking and the application of public policy, it is normal, and even encouraged, for rules to be left open to interpretation, with details filled by human judgment emerging from disputes in specific cases that are resolved after-the-fact. We offer

---

200 The Supreme Court, 2008 Term--Leading Cases, supra note 188, at 290.
201 Ricci, 557 U.S. at 585.
202 Id.
recommendations for dealing with this apparent mismatch, arguing for greater collaboration between experts in two different fields.

We emphasize that computer scientists cannot assume that the policy process will give them a meaningful, universal, and self-consistent theory of fairness to use as a specification for algorithms. There are structural, political, and jurisprudential reasons why no such theory exists today. Likewise, the policy process would likely not accept such a theory if it were generated by computer scientists.

At the same time, lawmakers and policymakers will need to adapt in light of these new technologies. We highlight changes that stem from automated decisionmaking. First, choices made when designing computer systems embed specific policy decisions and values in those systems whether or not they provide for accountability. Algorithms can, nevertheless, permit direct accountability to the public or to other third parties, despite the fact that full transparency is neither sufficient nor always necessary for accountability. For both groups, we note that the interplay between these areas will raise new questions and may generate new insights into what the goals of these decisionmaking processes should be.

A. Recommendations for Computer Scientists: Design for After-the-Fact Oversight

Computer scientists may tend to think of accountability in terms of compliance with a detailed specification set forth before the creation of an algorithm. For example, it is typical for programmers to define bugs based on the specification for a program--anything that differs from the specification is a bug; anything that follows it is a feature.203

This Section is intended to inform computer scientists that no one will remove all the ambiguities and offer them a clear, complete specification. Although lawmakers and policymakers can offer clarifications or other changes to guide the work done by developers,204 drafters may be unable to remove certain ambiguities for political reasons or be unwilling to resolve details to meet flexibility objectives. As such, computer scientists must account for the lack of precision--and the corresponding need for after-the-fact oversight by courts or other reviewers--when designing decisionmaking algorithms.

A computer scientist’s mindset can conflict deeply with many sources of authority to which developers may be responsible. Public opinion and social norms are inherently not precisely specified. The corporate requirements to satisfy one’s supervisor (or one’s supervisor’s supervisor) may not be clear. Perhaps most importantly and least intuitively for computer scientists, the operations of U.S. law


204 See infra Section IV.B.
and public policy also work against clear specifications. These processes often deliberately create ambiguous laws and guidance, leaving details—or sometimes even major concepts—open to interpretation.\textsuperscript{205} One cause of this ambiguity is the political reality of legislation. Legislators may be unable gather majority support to agree on the details of a proposed law, but may be able to get a majority of votes to pass relatively vague language that leaves various terms and conditions unspecified.\textsuperscript{206} For example, different legislators may support conflicting specific proposals that can be encompassed by a more general bill.\textsuperscript{207} Even legislators who do not know precisely what they want may still object to a particular proposed detail; each detail that caused sufficient objections would need to be stripped out of a bill before it could become law.\textsuperscript{208} Another explanation of ambiguity is that legislators may have uncertainty about the situations to which a law or policy will apply. Drafters may worry that they have not fully considered all of the possibilities. This creates an incentive to build in enough flexibility to cover unexpected circumstances that currently exist or may exist in the future.\textsuperscript{209} The U.S. Constitution is often held up as a model in this regard: generalized provisions for governance and individual rights continue to be applicable even as the landscape of society changes dramatically.\textsuperscript{210} Finally, ambiguity may stem from shared uncertainty about how best to solve even a known problem. Here, drafters may feel that they know what situations will arise but still not know how they want to deal with them. They may, in effect, choose to delegate authority to other parties by underspecifying particular aspects of a law or policy. Vagueness supports experimentation to help determine what methods are most effective or desirable.\textsuperscript{211}

\begin{flushleft}
\textsuperscript{205} See, e.g., Marbury v. Madison, 5 U.S. (1 Cranch) 137 (1803) (establishing the practice of judicial review, on which the Constitution was silent); 47 U.S.C. § 222(c)(1) (2012) (requiring a telecommunications carrier to get the “approval of the customer” to use or disclose customer proprietary network information, but neglecting to define “approval”).
\textsuperscript{206} See Victoria F. Nourse & Jane S. Schacter, The Politics of Legislative Drafting: A Congressional Case Study, 77 N.Y.U. L. Rev. 575, 593 (2002) (“Several staffers thought that pressures of time, and the political imperative to get a bill ‘done,’ bred ambiguity. Indeed, one staffer emphasized that while it was well and good to draft a bill clearly, there was no guarantee that the clear language would be passed by the House or make it through conference.”).
\textsuperscript{208} Id. at 1.
\textsuperscript{209} See, e.g., 17 U.S.C. § 1201 (2012) (granting the Copyright Office the power to create exemptions from the statute’s prohibition on anti-circumvention).
\textsuperscript{210} See David A. Strauss, The Living Constitution (2010). Laws governing law enforcement access to personal electronic records are often cited as a counterexample, with over-specific provisions in the Electronic Communications Privacy Act (18 U.S.C. §§ 2510-2704 (2012)) that fail to account for a shift in technology to a regime where most records reside with third party service providers, not users’ own computers. For a more detailed explanation, see Orin S. Kerr, Applying the Fourth Amendment to the Internet: A General Approach, 62 Stan. L. Rev. 1005 (2010).
\textsuperscript{211} A similar logic—policy experimentation among the states is one of the principles underlying federalism. See New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (praising the ability of a state to “serve as a laboratory” for democracy).
\end{flushleft}
The United States has a long history of dealing with these ambiguities through after-the-fact and retroactive oversight by the courts. In our common law system, ambiguities and uncertainties are left unaddressed until there is a dispute and their resolution becomes necessary. Disagreements about the application of a law or regulation to a specific set of facts are resolved through cases, and the areas of ambiguity are clarified over time by the accretion of many rulings on specific situations. Even when statutes and regulations may have specific and detailed language, they are interpreted through cases—with extensive deference often given to the expertise of administrative agencies. Those cases form binding precedents, which, in the U.S. common law system, are an additional source of legal authority alongside the statutes themselves. The gradual development and extension of law and regulations through cases with specific fact patterns allows for careful consideration of meaning and effects at a level of granularity that is usually impossible to reach during the drafting process.

In practice, these characteristics imply that computer scientists should focus on creating algorithms that are reviewable, not just compliant with the specifications that are generated in the drafting process. For example, this means it would have been good for the Diversity Visa Lottery described in Section II.C to use an algorithm that made fair, random choices and it would be desirable for the State Department to be able to demonstrate that property to a court or a skeptical lottery participant. The technical approaches described in this Article provide several ways for algorithm designers to ensure that the actual basis for a decision can be verified later. With these tools, reviewers can check whether an algorithm actually was used to make a particular decision, whether random inputs were chosen fairly, and whether the algorithm comports with certain principles specified at the time of the design. Essentially, these technical tools allow continued after-the-fact evaluations of algorithms by allowing for and assisting the judicial system’s

212 See generally E. Allan Farnsworth, An Introduction to the Legal System of the United States (Steve Sheppard ed., 4th ed. 2010).
213 See generally id.
215 See Farnsworth, supra note 213.
216 Id.
217 Another possible conclusion is that certain algorithms should also be developed to be flexible, permitting adaptation as new cases, laws, or regulations add to the initial specifications. The need to adapt algorithms is discussed further in subsection IV.B.1. This also reflects the current insufficiency of building a system in accord with a particular specification, though oversight or enforcement bodies evaluating the decision at a later point in time will still need to be able to certify compliance with any actual specifications.
218 Algorithms offer a new opportunity for decisionmaking processes to be reviewed by nongovernmental organizations. We discuss this possibility further in subsection IV.B.2.
219 See supra Sections II.B & III.B.
220 See supra Section II.C.
221 See supra notes XX-YY and accompanying text.
222 See supra notes XX-YY and accompanying text.
traditional role in ultimately determining the legality of particular decisionmaking.223

Implementing the approaches described in this Article would improve the accountability of decisionmaking algorithms dramatically, but we see that implementation as only a first step. We encourage research into extensions of these technical tools, as well as new techniques designed to facilitate oversight.

B. Recommendations for Lawmakers and Policymakers

The other side of the coin is that lawmakers and policymakers need to recognize and adapt to the changes wrought by algorithmic decisionmaking. Characteristics of algorithms offer both new opportunities and new challenges for the development of legal regimes governing decisionmaking: algorithmic decisionmaking can reduce the benefits of ambiguity, increase accountability to the public, and permit greater accountability than was previously possible in cases where aspects of the decision process remain secret.

1. Reduced Benefits of Ambiguity

Although computer scientists can build algorithms to permit after-the-fact assessment and accountability, they cannot alter the fact that any algorithm design will encode specific values and involve specific rules. Furthermore, the design of a computer system may limit opportunities for after-the-fact accountability. In other words, if a system is not designed to permit certification of a particular characteristic, an oversight body cannot be certain that it will be able to certify that characteristic. Both of these traits imply that automated decisionmaking can exacerbate certain disadvantages of legal ambiguities.

In the framework set forth above,224 we identify key drivers for ambiguity: political stalemate, uncertainty about future circumstances, and desire for policy experimentation. Here, with respect to each of these drivers, we will discuss how the shift to algorithmic decisionmaking diminishes the appeal of ambiguity, and we will suggest ways of retaining the functional benefits that ambiguity provides in the U.S. lawmaking system and ways that are more amenable to automation.

Ambiguity stemming from political stalemate essentially passes the buck for determining details from legislators to someone later on in the process. These later actors tend to be more sheltered from political pressures and thus able to make specific decisions without risking their jobs at the next election. Judges and administrative agencies frequently fill this role. Courts are expected to offer impartial decisions resistant to public pressure225 and administrative agencies are

---

223 Computer scientists model this after-the-fact input as an “oracle” that can be consulted only rarely on the acceptability of the algorithm. See Kroll, supra note 119.

224 Supra Section IV.A.

225 See, e.g., The Federalist No. 78 (Alexander Hamilton), (laying out the philosophy that the judiciary’s role is to secure an “impartial administration of the laws”). However, the rise of elected judges raises questions about this traditional role of the court system. See Stephen J. Choi et al., Professionals or Politicians: The Uncertain Empirical Case for an Elected Rather Than Appointed Judiciary (Univ. of Chi. Law Sch., John M. Olin Law & Economics Working Paper No. 357, 2007) (finding that elected judges behave more like politicians than appointed independent judges).
expected to retain staff who offer subject matter expertise beyond what is expected of legislators, despite changes in political administrations.\(^{226}\)

However, this transfer of responsibility often works in less than ideal ways when it comes to software systems.\(^{227}\) Fully automated decisionmaking may exacerbate these problems by adding another actor to whom the responsibility can devolve: the developer who programs the decisionmaking software. Citron offers examples of failures in automated systems that determine benefits eligibility, notably the airport “No Fly” lists, terrorist identifications, and punishment for “dead-beat” parents.\(^{228}\) Lawmakers should consider this possibility and avoid giving the responsibility for filling in the details of the law to program developers because (1) the algorithms will apply broadly, affecting all participants; (2) the program developer is unlikely to be held accountable by the current political process; and (3) the program developer is unlikely to have substantive expertise about the political decision being made.\(^{229}\)

One potential method for restricting the discretion of developers without requiring specifications in the legislation itself would be for administrative agencies to publish guidance for software development. Difficulties in translating between code choices and policy effects still would exist, but they could be partly eased using the technical methods we have described.\(^{230}\) For example, administrative agencies could work together with developers to identify the properties they want a piece of software to possess, and the program could then be designed to satisfy those properties and permit proof.

Ambiguity generated by uncertainty about the situational circumstances or ambiguity motivated by a desire for policy experimentation presents a more difficult concern. Here, the problem raised by automated decisionmaking is that a piece of software locks in a particular interpretation of law for the duration of its use, and, especially in government contexts, provisions to update the software code may not be made. Worries about changing or unexpected circumstances could be assuaged by adding sunset provisions to software systems,\(^{231}\) requiring periodic


\(^{227}\) For example, Citron argues that “[d]istortions in policy have been attributed to the fact that programmers lack ‘policy knowledge.’” and that this leads to software that does not reflect policy goals. Citron, supra note 6, at 1261. Ohm also reports on a comment of Felten that “[i]n technology policy debates, lawyers put too much faith in technical solutions, while technologists put too much faith in legal solutions.” Paul Ohm, Breaking Felten’s Third Law: How Not to Fix the Internet, 87 Denver L. Rev. Online (2010), http://www.denverlawreview.org/how-to-regulate/2010/2/22/breaking-feltens-third-law-how-not-to-fix-the-internet.html [https://perma.cc/6RGQ-KUMW] (internal quotation marks omitted).

\(^{228}\) Citron, supra note 6, at 1256-57.

\(^{229}\) Id. at 6. A distinction should be drawn here between the responsibilities given to individual developers of particular algorithms and the responsibilities given to computer scientists in general. Great gains can be made by improved dialogue between computer scientists and lawmakers and policymakers about how to design algorithms to reach social goals.

\(^{230}\) See supra Sections II.B & III.B.

\(^{231}\) The effectiveness of sunset provisions in leading to actual reconsideration and change is debatable. The inertia of the pre-existing choices can be hard to overcome. See, e.g., Mark A.
review and reconsideration of the software. Additionally, software should be designed with eventual revisions and updates in mind. As for preserving the benefits of policy experimentation, the traditional solution might be having multiple programs that take multiple approaches deployed simultaneously. A more sophisticated version of this solution is the incorporation of machine learning into decisionmaking systems. Again, machine learning can have its own fairness pitfalls, and care should be taken to consider fair machine learning methods and to build in precautions like persistent testing of the hypotheses built into the machine learning model.

More generally, the benefits of ambiguity decrease in the case of algorithmic decisionmaking. Here, an uninformed programming actor may determine the details and then apply them broadly. In addition, the choice of algorithm cements the particular policy choices encoded in that software for as long as it is used. Drafters should instead consider whether they should increase the specificity offered by law and policy governing these algorithms to prevent coders from filling the ambiguity.

To a certain extent, this question mirrors the rules versus standards debate about the relative merits of laws that specify actions and their repercussions (for example, a speed limit) and those that espouse a principle open to interpretation (for example, “drive at a speed reasonable for the conditions”). Rules give clarity and forewarning, while standards offer greater flexibility for interpretation.

Here, the question is whether drafters should include additional and clearer specifications for developers. In practice, drafters may wish to incorporate a set of narrow rules within a broad, overarching standard. For example, drafters could include specifications of each of the properties that they want a piece of software to possess and requirements that the developer design that program in a way that renders those properties provable upon review. Additionally, drafters might consider requiring a general statement of purpose for the algorithm. Doing so would give the developer some flexibility in writing the code while also ensuring that particular properties can be checked later.

232 See supra note XX (noting that machine learning programs give predictions but not confidence levels).
233 See supra Sections III.A-B.
234 In other words, even after a machine learning algorithm determines that a particular rule should be used to produce particular results, it always should continue to test inputs that do not follow that rule. See, e.g., Russell & Norvig, supra note 68.
235 See, e.g., Louis Kaplow, Rules versus Standards: An Economic Analysis, 42 Duke L.J. 557, 562-66 (1992) (arguing that rules are more costly to promulgate while standards are most costly on individuals).
2. Accountability to the Public

Oversight is traditionally performed by courts, enforcement agencies, or other designated entities such as government prosecutors. Typically, the public and third parties have an indirect oversight role through the ability to provide political feedback and the ability to bring lawsuits if their specific circumstances allow. The use of software can alter how effectively the legal system and the public can oversee the decisionmaking process.

In one sense, decisionmaking computer systems can enhance accountability to the public and interested third parties by permitting greater involvement in oversight. The technical tools we describe allow for a more direct form of oversight by these parties. Unlike traditional legal oversight mechanisms that generally require discovery or the gathering of internal evidence, the technical tools may enable verifications by the public and by third parties that are completely independent from the organizations using the software. For example, technologically proficient members of the public or third parties could verify that a particular algorithm was used in a program or that the program has particular properties. In addition, a system could be built to enable participants to check these properties for their own outcomes so that nontechnical users could verify these facts while the system as a whole would be overseen by others--potentially both inside and outside of government--who have the necessary technological expertise. As another example, third parties could be involved in generating fair randomness.

In contrast to the possibility for enhanced public accountability, the use of software without the reliance on technical tools for oversight, as we have described, can reduce accountability to the public because courts and other policy actors are generally ill-equipped to evaluate software, thereby hampering our traditional scrutiny of decisionmaking. The U.S. court system is designed to protect against wrongful government actions through the power of judicial review. Judicial review gives judges the power and responsibility to determine if government actions comply with legal obligations. Similarly, for private actions, the legal system vests judges and regulatory agencies with the authority to determine whether those actions are consistent with legal standards.

The use of software systems to make decisions shifts these burdens to external experts or to the organizations creating and deploying the software. Courts and enforcement agencies are no longer able to make a determination as to whether the rules have been properly applied or whether fairness obligations have been met. That determination shifts to the experts evaluating the automated decisionmaking process. One way to address this unintended shift in responsibility is to appoint

---

237 See Farnsworth, supra note 213.
238 The public can vote political leaders out of office and aggrieved parties can bring lawsuits to seek vindication.
239 See supra note 117 (using a quantum source to generate randomness).
240 See Marbury v. Madison, 5 U.S. (1 Cranch) 137, 177 (1803) (“It is emphatically the province and duty of the judicial department to say what the law is.”).
technical experts as special masters. Courts typically appoint special masters to perform functions on behalf of the court that require special skill or knowledge.\(^{241}\)

Another issue that challenges public accountability is the validation of the technical tools we have described. For courts, technical tools cannot be accepted until their integrity and reliability are proven. Courts have long confronted the problem of the admissibility of scientific evidence. There is a rich literature about the standards courts should use to admit expert scientific evidence, and, even now, federal and state standards vary.\(^{242}\) Following many years of debate, the Federal Rules of Evidence now spell out the federal standard for the acceptability of new scientific methods in adversarial proceedings.\(^{243}\) In 1993, the Supreme Court adjusted those standards to take account of factors that include testing, peer review and publication.\(^{244}\) The courts took years during the 1980s and 90s to establish and accept the scientific validity of DNA and the methods used to isolate and test DNA\(^{245}\) and even now, there are concerns that some scientific methods (e.g. matching DNA based on mixtures of DNA) may be receiving undeserved deference from courts and thus resulting in faulty findings of fact.\(^{246}\) The Federal Rules of Evidence now provide for the acceptability of new scientific methods in adversarial proceedings.\(^{247}\) In 1993, the Supreme Court set out standards to meet the Federal Rules requirements that include testing, peer review and publication.\(^{248}\) This addresses the validation of technical tools used to examine automated

\(^{241}\) See, e.g., United States v. Microsoft Corp., 147 F.3d 935, 959 n.4 (D.C. Cir. 1998) (noting Larry Lessig’s role as a special master for technical issues in the antitrust case brought against Microsoft).


\(^{243}\) See Fed. R. Evid. 702 (“A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if . . . the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue”).

\(^{244}\) See Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 592-95 (1993) (explaining that a judge faced with a proffer of expert scientific testimony must assess whether the testimony’s underlying reasoning is valid, and in doing so, consider whether the technique or theory in question can be tested and whether it has been subjected to peer review and publication).

\(^{245}\) See, e.g., Nat’l Research Council, The Evaluation of Forensic DNA Evidence 166-211 (1996) (discussing the legal implications of the use of forensic DNA testing as well as the procedural and evidentiary rules that affect such use).


\(^{247}\) See Fed. R. Evid. 702 (“A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if . . . the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue”).

\(^{248}\) See Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 592-95 (1993) (explaining that a judge faced with a proffer of expert scientific testimony must assess whether the testimony’s underlying reasoning is valid, and, in doing so, consider whether the technique or theory in question can be tested and whether it has been subjected to peer review and publication).
decisionmaking, but still leaves open the assurance of the technical tools’ reliability. Ordinarily, the U.S. legal system relies on the adversarial process to assure the accuracy of findings. This attribute may be preserved by allowing multiple experts to test software-driven processes.

3. Secrets and Accountability

Implementing automated decisionmaking in a socially and politically acceptable way requires progress in our ability to communicate and understand fine-grained partial information about how decisions are reached. Full transparency (disclosing everything) is technically trivial but politically and practically infeasible and may not be useful, as described in Section II.A. However, disclosing nothing about the basis for a decision is socially unacceptable and generally poses a technical challenge. Lawmakers and policymakers should remember that it is possible to make an algorithm accountable without the evaluator having full access to the algorithm.249

U.S. law and policy often focus on transparency and sometimes even equate oversight with transparency for the overseer.250 As such, accountability without full transparency may seem counterintuitive. However, oversight based on partial information occurs regularly within the legal system. Courts prevent consideration of many types of information for various policy reasons: disclosures of classified information may be prevented or limited to preserve national security;251 juvenile records may be sealed because of the notion that mistakes made in one’s youth should not follow them forever;252 and other evidence is deemed inadmissible for a multitude of reasons, including being unscientific,253 hearsay,254 inflammatory,255 or illegally obtained.256 Thus, all of the rules of evidence could be construed as precedent for the idea that optimal oversight does not require full information.

There are strong policy justifications for holding back information in the case of automated decisionmaking. Revealing software source code and input data can expose trade secrets, violate privacy, hamper law enforcement, or lead to gaming of the decisionmaking process.257 The advantage of computer systems is that concealment of code and data does not imply an inability to analyze the code and data. The technical tools we describe give lawmakers and policymakers the ability to keep software programs and their inputs secret while still rendering them

249 See supra note XXX.
251 See 18 U.S.C. § 798(a) (2012) (providing that the disclosure of classified government information may result in criminal liability).
252 See, e.g., N.Y. Crim. Proc. § 720.15 (requiring filing under seal in juvenile proceedings).
253 See Fed. R. Evid. 702 (establishing the court’s discretion to admit scientific evidence).
254 See Fed. R. Evid. 802 (stating that hearsay evidence is inadmissible unless a federal statute, the rules of evidence, or the Supreme Court provides otherwise).
255 See Fed. R. Evid. 403 (providing for the exclusion of relevant evidence for prejudice).
257 See Section II.A.
accountable. They can implement these tools in government-run algorithms, such as the DVL, and incentivize nongovernmental actors to use them, perhaps by mandating use or by requiring transparency—at least to courts—of code and inputs if they do not employ such technical tools.
Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm’s use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

This Essay examines these concerns through the lens of American antidiscrimination law—more particularly, through Title
VII’s prohibition of discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining’s victims would seem to lie in disparate impact doctrine. Case law and the Equal Employment Opportunity Commission’s Uniform Guidelines, though, hold that a practice can be justified as a business necessity when its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations. Unless there is a reasonably practical way to demonstrate that these discoveries are spurious, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others’ discrimination against members of protected groups, or flaws in the underlying data.

Addressing the sources of this unintentional discrimination and remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically. There are a number of practical limits to what can be accomplished computationally. For example, when discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical data to rid it of this taint. Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain. These challenges for reform throw into stark relief the tension between the two major theories underlying antidiscrimination law: anticlassification and antisubordination. Finding a solution to big data’s disparate impact will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of “discrimination” and “fairness.”

Introduction ......................................................................................................................... 673
I. How Data Mining Discriminates .......................................................................................... 677
   A. Defining the “Target Variable” and “Class Labels” ..................................................... 677
   B. Training Data .............................................................................................................. 680
      1. Labeling Examples .................................................................................................. 681
      2. Data Collection ...................................................................................................... 684
   C. Feature Selection ...................................................................................................... 688
   D. Proxies ...................................................................................................................... 691
   E. Masking ..................................................................................................................... 692
II. Title VII Liability for Discriminatory Data Mining .......................................................... 694
   A. Disparate Treatment ................................................................................................. 694
   B. Disparate Impact ....................................................................................................... 701
   C. Masking and Problems of Proof ............................................................................. 712
III. The Difficulty for Reforms ............................................................................................ 714
   A. Internal Difficulties ................................................................................................. 715
INTRODUCTION

“Big Data” is the buzzword of the decade. Advertisers want data to reach profitable consumers, medical professionals to find side effects of prescription drugs, supply-chain operators to optimize their delivery routes, police to determine where to focus resources, and social scientists to study human interactions. Though useful, however, data is not a panacea. Where data is used predictively to assist decision making, it can affect the fortunes of whole classes of people in consistently unfavorable ways. Sorting and selecting for the best or most profitable candidates means generating a model with winners and losers. If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.

Although we live in the post–civil rights era, discrimination persists in American society and is stubbornly pervasive in employment, housing, credit, and consumer markets. While discrimination certainly endures in part due to decision makers’ prejudices, a great deal of modern-day inequality can be attributed to what sociologists call “institutional” discrimination. Unconscious, implicit biases and inertia within society’s institutions, rather than intentional

8. Id.
choices, account for a large part of the disparate effects observed.\textsuperscript{9} Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society. It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment.

Algorithms\textsuperscript{10} could exhibit these tendencies even if they have not been manually programmed to do so, whether on purpose or by accident. Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight. Such a possibility has gone unrecognized by most scholars and policy makers, who tend to fear concealed, nefarious intentions or the overlooked effects of human bias or error in hand coding algorithms.\textsuperscript{11} Because the discrimination at issue is unintentional, even honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions. Furthermore, because the mechanism through which data mining may disadvantage protected classes is less obvious in cases of unintentional discrimination, the injustice may be harder to identify and address.

In May 2014, the White House released a report titled Big Data: Seizing Opportunities, Preserving Values (Podesta Report), which hinted at the discriminatory potential of big data.\textsuperscript{12} The report finds “that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”\textsuperscript{13} It suggests that there may be unintended discriminatory

\begin{flushright}
\textsuperscript{9} See Andrew Grant-Thomas & John a. powell, Toward a Structural Racism Framework, 15 POVERTY & RACE 3, 4 (“‘Institutional racism’ was the designation given in the late 1960s to the recognition that, at very least, racism need not be individualist, essentialist or intentional.”).

\textsuperscript{10} An “algorithm” is a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions. SOLON BAROCAS ET AL., DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER (2014), http://www.datacivilrights.org/pubs/2014-1030/Technology.pdf [https://perma.cc/X3YX-XHNA]. Algorithms play a role in both automating the discovery of useful patterns in datasets and automating decision making that relies on these discoveries. This Essay uses the term to refer to the latter.


\textsuperscript{13} Id. (introductory letter).
\end{flushright}
effects from data mining but does not detail how they might come about. Because the origin of the discriminatory effects remains unexplored, the report’s approach does not address the full scope of the problem.

The Podesta Report, as one might expect from the executive branch, seeks to address these effects primarily by finding new ways to enforce existing law. Regarding discrimination, the report primarily recommends that enforcement agencies, such as the Department of Justice, Federal Trade Commission, Consumer Financial Protection Bureau, and Equal Employment Opportunity Commission (EEOC), increase their technical expertise and “develop a plan for investigating and resolving violations of law in such cases.”

As this Essay demonstrates, however, existing law largely fails to address the discrimination that can result from data mining. The argument is grounded in Title VII because, of all American antidiscrimination jurisprudence, Title VII has a particularly well-developed set of case law and scholarship. Further, there exists a rapidly emerging field of “work-force science,” for which Title VII will be the primary vehicle for regulation. Under Title VII, it turns out that some, if not most, instances of discriminatory data mining will not generate liability. While the Essay does not show this to be true outside of Title VII itself, the problem is likely not particular to Title VII. Rather, it is a feature of our current approach to antidiscrimination jurisprudence, with its focus on procedural fairness. The analysis will likely apply to other traditional areas of discrimination, such as housing or disability discrimination. Similar tendencies to disadvantage the disadvantaged will likely arise in areas that regulate legitimate economic discrimination, such as credit and insurance.

This Essay proceeds in three Parts. Part I introduces the computer science literature and proceeds through the various steps of solving a problem with data mining: defining the target variable, labeling and collecting the training data, using feature selection, and making decisions on the basis of the resulting model. Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes. Finally, Part I notes that data mining poses the additional problem of

14. Id. at 64 (“This combination of circumstances and technology raises difficult questions about how to ensure that discriminatory effects resulting from automated decision processes, whether intended or not, can be detected, measured, and redressed.”).
15. Id. at 65.
giving data miners the ability to disguise intentional discrimination as accidental.

In Part II, the Essay reviews Title VII jurisprudence as it applies to data mining. Part II discusses both disparate treatment and disparate impact, examining which of the various data mining mechanisms identified in Part I will trigger liability under either Title VII theory. At first blush, either theory is viable. Disparate treatment is viable because data mining systems treat everyone differently; that is their purpose. Disparate impact is also viable because data mining can have various discriminatory effects, even without intent. But as Part II demonstrates, data mining combines some well-known problems in discrimination doctrines with new challenges particular to data mining systems, such that liability for discriminatory data mining will be hard to find. Part II concludes with a discussion of the new problems of proof that arise for intentional discrimination in this context.

Finally, Part III addresses the difficulties reformers would face in addressing the deficiencies found in Part II. These difficulties take two forms: complications internal to the logic of data mining and political and constitutional difficulties external to the problem. Internally, the different steps in a data mining problem require constant subjective and fact-bound judgments, which do not lend themselves to general legislative resolution. Worse, many of these are normative judgments in disguise, about which there is not likely to be consensus. Externally, data mining will force society to explicitly rebalance the two justifications for antidiscrimination law—rooting out intentional discrimination and equalizing the status of historically disadvantaged communities. This is because methods of proof and corrective measures will often require an explicit commitment to substantive remediation rather than merely procedural remedies. In certain cases, data mining will make it simply impossible to rectify discriminatory results without engaging with the question of what level of substantive inequality is proper or acceptable in a given context. Given current political realities and trends in constitutional doctrines, legislation enacting a remedy that results from these discussions faces an uphill battle. To be sure, data mining also has the potential to help reduce discrimination by forcing decisions onto a more reliable empirical foundation and by formalizing decision-making processes, thus limiting the opportunity for individual bias to affect important assessments. In many situations, the introduction of data mining will be a boon to civil rights, even where it fails to root out discrimination altogether, and such efforts should be encouraged. Yet, understanding when and why discrimination persists in cases of data-driven decision making reveals important and sometimes troubling limits to the promise of big data, for which there are no ready solutions.

I. HOW DATA MINING DISCRIMINATES

Although commentators have ascribed myriad forms of discrimination to data mining, there remains significant confusion over the precise mechanisms that render data mining discriminatory. This Part develops a taxonomy that isolates and explicates the specific technical issues that can give rise to models whose use in decision making may have a disproportionately adverse impact on protected classes. By definition, data mining is always a form of statistical (and therefore seemingly rational) discrimination. Indeed, the very point of data mining is to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar. Nevertheless, data mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage. Unlike more subjective forms of decision making, data mining’s ill effects are often not traceable to human bias, conscious or unconscious. This Part describes five mechanisms by which these disproportionately adverse outcomes might occur, walking through a sequence of key steps in the overall data mining process.

A. Defining the “Target Variable” and “Class Labels”

In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset. In particular, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision making can rely. The accumulated set of discovered relationships is commonly called a “model,” and these models can be employed to automate the process of classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes. Familiar examples of such applications include spam or fraud detection, credit scoring, and insurance pricing. These examples all involve attempts to determine the status or likely outcome of cases under consideration based solely on access to correlated data. Data mining helps identify cases of
 spam and fraud and anticipate default and poor health by treating these states and outcomes as a function of some other set of observed characteristics. In particular, by exposing so-called “machine learning” algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm “learns” which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest.

Two concepts from the machine learning and data mining literature are important here: “target variables” and “class labels.” The outcomes of interest discussed above are known as target variables. While the target variable defines what data miners are looking for, “class labels” divide all possible values of the target variable into mutually exclusive categories.

The proper specification of the target variable is frequently not obvious, and the data miner’s task is to define it. To start, data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse. In particular, data miners must determine how to solve the problem at hand by translating it into a question about the value of some target variable. The open-endedness that characterizes this part of the process is often described as the “art” of data mining. This initial step requires a data miner to “understand[] the project objectives and requirements from a business perspective [and] then convert[] this knowledge into a data mining problem definition.” Through this necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes.

Problem specification is not a wholly arbitrary process, however. Data mining can only address problems that lend themselves to formalization as questions about the state or value of the target variable. Data mining works exceedingly well for dealing with fraud and spam because these cases rely on extant, binary categories. A given instance either is or is not fraud or spam, and the definitions of fraud or spam are, for the most part, uncontroversial. A computer can then flag or refuse transactions or redirect emails according to

---

22. Id.
23. Id.
24. Comm. on the Analysis of Massive Data et al., Frontiers in Massive Data Analysis 101 (2013), http://www.nap.edu/catalog.php?record_id=18374 [https://perma.cc/5DNQ-UFE4]. The machine learning community refers to classification, estimation, and prediction—the techniques that we discuss in this Essay—as “supervised” learning because analysts must actively specify a target variable of interest. Id. at 104. Other techniques known as “unsupervised” learning do not require any such target variables and instead search for general structures in the dataset, rather than patterns specifically related to some state or outcome. Id. at 102. Clustering is the most common example of “unsupervised” learning, in that clustering algorithms simply reveal apparent hot spots when plotting the data in some fashion. Id. We limit the discussion to supervised learning because we are primarily concerned with the sorting, ranking, and predictions enabled by data mining.
well-understood distinctions. In these cases, data miners can simply rely on these simple, preexisting categories to define the class labels.

Sometimes, though, defining the target variable involves the creation of new classes. Consider credit scoring, for instance. Although now taken for granted, the predicted likelihood of missing a certain number of loan repayments is not a self-evident answer to the question of how to successfully extend credit to consumers. Unlike fraud or spam, “creditworthiness” is an artifact of the problem definition itself. There is no way to directly measure creditworthiness because the very notion of creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system. That is, an individual’s ability to repay some minimum amount of an outstanding debt on a monthly basis is taken to be a nonarbitrary standard by which to determine in advance and all-at-once whether he is worthy of credit.

Data mining has many uses beyond spam detection, fraud detection, credit scoring, and insurance pricing. As discussed in the introduction, this Essay will focus on the use of data mining in employment decisions. Extending this discussion to employment, then, where employers turn to data mining to develop ways of improving and automating their search for good employees, they face a number of crucial choices.

Like creditworthiness, the definition of a good employee is not a given. “Good” must be defined in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example. When employers mine data for good employees, they are, in fact, looking for employees whose observable characteristics suggest that they would meet or exceed some monthly sales threshold, perform some task in less than a certain amount of time, or remain in their positions for more than a set number of weeks or months. Rather than drawing categorical distinctions along these lines, data mining could also estimate or predict the specific numerical value of sales, production time, or tenure period, enabling employers to rank rather than simply sort employees.

These may seem like eminently reasonable things for employers to want to predict, but they are, by necessity, only part of an array of possible definitions of “good.” An employer may instead attempt to define the target variable in a more holistic way—by, for example, relying on the grades that prior employees have received in annual reviews, which are supposed to reflect

---

27. Though described as a matter of detection, this is really a classification task, where any given transaction or email can belong to one of two possible classes, respectively: fraud or not fraud, or spam or not spam.


an overall assessment of performance. These target variable definitions simply inherit the formalizations involved in preexisting assessment mechanisms, which in the case of human-graded performance reviews, may be far less consistent.  

Thus, the definition of the target variable and its associated class labels will determine what data mining happens to find. While critics of data mining have tended to focus on inaccurate classifications (false positives and false negatives), as much—if not more—danger resides in the definition of the class label itself and the subsequent labeling of examples from which rules are inferred. While different choices for the target variable and class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different choices may have a greater or lesser adverse impact on protected classes. For example, as later Parts will explain in detail, hiring decisions made on the basis of predicted tenure are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity. If the turnover rate happens to be systematically higher among members of certain protected classes, hiring decisions based on predicted length of employment will result in fewer job opportunities for members of these groups, even if they would have performed as well as or better than the other applicants the company chooses to hire.

B. Training Data

As described above, data mining learns by example. Accordingly, what a model learns depends on the examples to which it has been exposed. The data that function as examples are known as “training data”—quite literally, the data that train the model to behave in a certain way. The character of the training data can have meaningful consequences for the lessons that data mining happens to learn. As computer science scholars explain, biased training data leads to discriminatory models. This can mean two rather different things,

30. Joseph M. Stauffer & M. Ronald Buckley, The Existence and Nature of Racial Bias in Supervisory Ratings, 90 J. APPLIED PSYCHOL. 586, 588–89 (2005) (showing evidence of racial bias in performance evaluations). Nevertheless, devising new target variables can have the salutary effect of forcing decision makers to think much more concretely about the outcomes that justifiably determine whether someone is a “good” employee. The explicit enumeration demanded of data mining thus also presents an opportunity to make decision making more consistent, more accountable, and fairer overall. This, however, requires conscious effort and careful thinking, and is not a natural consequence of adopting data mining.


32. See infra Part I.B.

though: (1) if data mining treats cases in which prejudice has played some role as valid examples to learn from, that rule may simply reproduce the prejudice involved in these earlier cases; or (2) if data mining draws inferences from a biased sample of the population, any decision that rests on these inferences may systematically disadvantage those who are under- or overrepresented in the dataset. Both can affect the training data in ways that lead to discrimination, but the mechanisms—improper labeling of examples and biased data collections—are sufficiently distinct that they warrant separate treatment.

1. Labeling Examples

Labeling examples is the process by which the training data is manually assigned class labels. In cases of fraud or spam, the data miners draw from examples that come prelabeled: when individual customers report fraudulent charges or mark a message as spam, they are actually labeling transactions and email for the providers of credit and webmail. Likewise, an employer using grades previously given at performance reviews is also using prelabeled examples.

In certain cases, however, there may not be any labeled data and data miners may have to figure out a way to label examples themselves. This can be a laborious process, and it is frequently fraught with peril.34 Often the best labels for different classifications will be open to debate. On which side of the creditworthy line does someone who has missed four credit card payments fall, for example?35 The answer is not obvious. Even where the class labels are uncontested or uncontroversial, they may present a problem because analysts will often face difficult choices in deciding which of the available labels best applies to a particular example. Certain cases may present some, but not all, criteria for inclusion in a particular class.36 The situation might also work in reverse, where the class labels are insufficiently precise to capture meaningful differences between cases. Such imperfect matches will demand that data miners exercise judgment.

The unavoidably subjective labeling of examples will skew the resulting findings such that any decisions taken on the basis of those findings will characterize all future cases along the same lines. This is true even if such

34. Hand, supra note 26, at 10–11.
35. Id. at 10 (“The classical supervised classification paradigm also takes as fundamental the fact that the classes are well defined. That is, that there is some fixed clear external criterion, which is used to produce the class labels. In many situations, however, this is not the case. In particular, when the classes are defined by thresholding a continuous variable, there is always the possibility that the defining threshold might be changed. Once again, this situation arises in consumer credit, where it is common to define a customer as ‘defaulting’ if they fall three months in arrears with repayments. This definition, however, is not a qualitative one (contrast has a tumor/does not have a tumor) but is very much a quantitative one. It is entirely reasonable that alternative definitions (e.g., four months in arrears) might be more useful if economic conditions were to change.”).
36. Id. at 11.
characterizations would seem plainly erroneous to analysts who looked more closely at the individual cases. For all their potential problems, though, the labels applied to the training data must serve as ground truth. Thus, decisions based on discoveries that rest on haphazardly labeled data or data labeled in a systematically, though unintentionally, biased manner will seem valid according to the customary validation methods employed by data miners. So long as prior decisions affected by some form of prejudice serve as examples of correctly rendered determinations, data mining will necessarily infer rules that exhibit the same prejudice.

Consider a real-world example from a different context as to how biased data labeling can skew results. St. George’s Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants based on its previous admissions decisions. Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants. In drawing rules from biased prior decisions, St. George’s Hospital unknowingly devised an automated process that possessed these very same prejudices. As editors at the British Medical Journal noted at the time, “[T]he program was not introducing new bias but merely reflecting that already in the system.” Were an employer to undertake a similar plan to automate its hiring decisions by inferring a rule from past decisions swayed by prejudice, the employer would likewise arrive at a decision procedure that simply reproduces the prejudice of prior decision makers. Indeed, automating the process in this way would turn the conscious prejudice or implicit bias of individuals involved in previous decision making into a formalized rule that would systematically alter the prospects of all future applicants. For example, the computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women’s or historically black colleges.

Not only can data mining inherit prior prejudice through the mislabeling of examples, it can also reflect current prejudice through the ongoing behavior of users taken as inputs to data mining. This is what Professor Latanya Sweeney discovered in a study that found that Google queries for black-sounding names were more likely to return contextual (i.e., key-word triggered)

37. Id. at 12. Even when evaluating a model, the kinds of subtle mischaracterizations that happen during training will be impossible to detect because most “evaluation data” is just a small subset of the training data that has been withheld during the learning process. Any problems with the training data will be present in the evaluation data.
39. Id. at 657.
40. Id.
advertisements for arrest records than those for white-sounding names. Sweeney confirmed that the companies paying for these advertisements had not set out to focus on black-sounding names; rather, the fact that black-sounding names were more likely to trigger such advertisements seemed to be an artifact of the algorithmic process that Google employs to determine which advertisements to display alongside certain queries. Although it is not fully known how Google computes the so-called “quality score” according to which it ranks advertisers’ bids, one important factor is the predicted likelihood, based on historical trends, that users will click on an advertisement. As Sweeney points out, the process “learns over time which [advertisement] text gets the most clicks from viewers [of the advertisement]” and promotes that advertisement in its rankings accordingly. Sweeney posits that this aspect of the process could result in the differential delivery of advertisements that reflect the kinds of prejudice held by those exposed to the advertisements. In attempting to cater to users’ preferences, Google will unintentionally reproduce the existing prejudices that inform users’ choices.

A similar situation could conceivably arise on websites that recommend potential employees to employers, as LinkedIn does through its Talent Match feature. If LinkedIn determines which candidates to recommend based on the demonstrated interest of employers in certain types of candidates, Talent Match will offer recommendations that reflect whatever biases employers happen to exhibit. In particular, if LinkedIn’s algorithm observes that employers disfavor certain candidates who are members of a protected class, Talent Match may decrease the rate at which it recommends these candidates to employers. The recommendation engine would learn to cater to the prejudicial preferences of employers.

There is an old adage in computer science: “garbage in, garbage out.” Because data mining relies on training data as ground truth, when those inputs

42. Id. at 48, 52. 
44. Sweeney, supra note 41, at 52.
45. The fact that black people may be convicted of crimes at a higher rate than nonblack people does not explain why those who search for black-sounding names would be any more likely to click on advertisements that mention an arrest record than those who see the same exact advertisement when they search for white-sounding names. If the advertisement implies, in both cases, that a person of that particular name has an arrest record, as Sweeney shows, the only reason the advertisements keyed to black-sounding names should receive greater attention is if searchers confer greater significance to the fact of prior arrests when the person happens to be black. Id. at 53.
are themselves skewed by bias or inattention, the resulting system will produce results that are at best unreliable and at worst discriminatory.

2. Data Collection

Decisions that depend on conclusions drawn from incorrect, partial, or nonrepresentative data may discriminate against protected classes. The individual records that a company maintains about a person might have serious mistakes,47 the records of the entire protected class of which this person is a member might also have similar mistakes at a higher rate than other groups, and the entire set of records may fail to reflect members of protected classes in accurate proportion to others.48 In other words, the quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records for certain classes of people). Even a dataset with individual records of consistently high quality can suffer from statistical biases that fail to represent different groups in accurate proportions. Much attention has focused on the harms that might befall individuals whose records in various commercial databases are error ridden.49 Far less consideration, however, has been paid to the systematic disadvantage that members of protected classes may suffer from being miscounted and, as a result, misrepresented in the evidence base.

Recent scholarship has begun to stress this point. Jonas Lerman, for example, worries about “the nonrandom, systemic omission of people who live on big data’s margins, whether due to poverty, geography, or lifestyle, and whose lives are less ‘datafied’ than the general population’s.”50 Professor Kate Crawford has likewise warned that “[b]ecause not all data is created or even collected equally, there are ‘signal problems’ in big-data sets—dark zones or shadows where some citizens and communities are overlooked or

47. Data quality is a topic of lively practical and philosophical debate. See, e.g., Luciano Floridi, Information Quality, 26 PHIL. & TECH. 1 (2013); Richard Y. Wang & Diane M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, 12 J. MGMT. INFO. SYS. 5 (1996). The components of data quality have been thought to include accuracy, precision, completeness, consistency, validity, and timeliness, though this catalog of features is far from settled. See generally LARRY P. ENGLISH, INFORMATION QUALITY APPLIED (2009).


49. See, e.g., FED. TRADE COMM’N, REPORT TO CONGRESS UNDER SECTION 319 OF THE FAIR AND ACCURATE CREDIT TRANSACTIONS ACT OF 2003 A-4 (2012) (finding that nearly 20 percent of consumers had an error in one or more of their three credit reports and that 5.4 percent of consumers had errors that could result in less favorable loan terms).

Errors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation. Not only will the quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data.

As an illustrative example, Crawford points to Street Bump, an application for Boston residents that takes advantage of accelerometers built into smart phones to detect when drivers ride over potholes. While Crawford praises the cleverness and cost-effectiveness of this passive approach to reporting road problems, she rightly warns that whatever information the city receives from Street Bump will be biased by the uneven distribution of smartphones across populations in different parts of the city. In particular, systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate. If the city were to rely on this data to determine where it should direct its resources, it would only further underserve these communities. Indeed, the city would discriminate against those who lack the capability to report problems as effectively as wealthier residents with smartphones.

A similar dynamic could easily apply in an employment context if members of protected classes are unable to report their interest in and qualification for jobs listed online as easily or effectively as others due to systematic differences in Internet access. The EEOC has established a program called “Eradicating Racism & Colorism from Employment” (E-RACE) that aims, at least in part, to prevent this sort of discrimination from occurring due

52. See id.; Lerman, supra note 50, at 57.
53. Crawford, supra note 51 (explaining that a sudden movement suggesting a broken road will automatically prompt the phone to report the location to the city).
54. Id.
55. See id.
56. This is, of course, a more general problem with representative democracy. For a host of reasons, the views and interests of the poor are relatively less well represented in the political process. See, e.g., Larry M. Bartels, Economic Inequality and Political Representation, in THE UNSUSTAINABLE AMERICAN STATE 167 (Lawrence Jacobs & Desmond King eds., 2009); MARTIN GILENS, AFFLUENCE AND INFLUENCE: ECONOMIC INEQUALITY AND POLITICAL POWER IN AMERICA (2012). The worry here, as expressed by Crawford, is that, for all its apparent promise, data mining may further obfuscate or legitimize these dynamics rather than overcome them.
to an employer’s desire for high-tech hiring, such as video résumés.\textsuperscript{57} E-RACE not only attempts to lower the barriers that would disproportionately burden applicants who belong to a protected class, but also ensures that employers do not develop an inaccurate impression of the incidence of qualified and interested candidates from these communities. If employers were to rely on tallies of high-tech candidates to direct their recruiting efforts, for example, any count affected by a reporting bias could have adverse consequences for specific populations systematically underrepresented in the dataset. Employers would deny equal attention to those who reside in areas incorrectly pegged as having a relatively lower concentration of qualified candidates.

Additional and even more severe risks may reside in the systematic omission of members of protected classes from such datasets. The Street Bump and Internet job application examples only discuss decisions that depend on raw tallies, rather than datasets from which decision makers want to draw generalizations and generate predictions. But data mining is especially sensitive to statistical bias because data mining helps to discover patterns that organizations tend to treat as generalizable findings even though the analyzed data only includes a partial sample from a circumscribed period. To ensure that data mining reveals patterns that hold true for more than the particular sample under analysis, the sample must be proportionally representative of the entire population, even though the sample, by definition, does not include every case.\textsuperscript{58}

If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or underrepresented class. While the representativeness of the data is often simply assumed, this assumption is rarely justified and is “perhaps more often incorrect than correct.”\textsuperscript{59} Data gathered for routine business purposes tend to lack the rigor of social scientific data collection.\textsuperscript{60} As Lerman points out, “Businesses may ignore or undervalue the preferences and behaviors of

\textsuperscript{57} Why Do We Need E-RACE?, EQUAL EMPLOY. OPPORTUNITY COMM’N, http://www1.eeoc.gov/eeoc/initiatives/e-race/why_e-race.cfm [https://perma.cc/S3GY-2MD6] (last visited Mar. 1, 2013). Due to the so-called “digital divide,” communities underserved by residential Internet access rely heavily on mobile phones for connectivity and thus often have trouble even uploading and updating traditional résumés. Kathryn Zickuhr & Aaron Smith, Digital Differences, PEW RES. CTR. (Apr. 13, 2012), http://www.pewinternet.org/2012/04/13/digital-differences [https://perma.cc/S545-42GY] (“Among smartphone owners, young adults, minorities, those with no college experience, and those with lower household income levels are more likely than other groups to say that their phone is their main source of internet access.”).

\textsuperscript{58} Data mining scholars have devised ways to address this known problem, but applying these techniques is far from trivial. See Sinno Jialin Pan & Qiang Yang, A Survey on Transfer Learning, 22 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENG’G 1345, 1354–56 (2010).

\textsuperscript{59} Hand, supra note 26, at 7.

consumers who do not shop in ways that big data tools can easily capture, aggregate, and analyze."\(^{61}\)

In the employment context, even where a company performs an analysis of the data from its entire population of employees—avoiding the apparent problem of even having to select a sample—the organization must assume that its future applicant pool will have the same degree of variance as its current employee base. An organization’s tendency, however, to perform such analyses in order to change the composition of their employee base should put the validity of this assumption into immediate doubt. The potential effect of this assumption is the future mistreatment of individuals predicted to behave in accordance with the skewed findings derived from the biased sample. Worse, these results may lead to decision procedures that limit the future contact an organization will have with specific groups, skewing still further the sample upon which subsequent analyses will be performed.\(^{62}\) Limiting contact with specific populations on the basis of unsound generalizations may deny members of these populations the opportunity to prove that they buck the apparent trend.

*Overrepresentation* in a dataset can also lead to disproportionately high adverse outcomes for members of protected classes. Consider an example from the workplace: managers may devote disproportionate attention to monitoring the activities of employees who belong to a protected class and consequently observe mistakes and transgressions at systematically higher rates than others, in part because these managers fail to subject others who behave similarly to the same degree of scrutiny. Not only does this provide managers with justification for their prejudicial suspicions, but it also generates evidence that overstates the relative incidence of offenses by members of these groups. Where subsequent managers who hold no such prejudicial suspicions cannot observe everyone equally, they may rely on this evidence to make predictions about where to focus their attention in the future and thus further increase the disproportionate scrutiny that they place on protected classes.

The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.

---

62. Practitioners, particularly those involved in credit scoring, are well aware that they do not know how the person purposefully passed over would have behaved if he had been given the opportunity. Practitioners have developed methods to correct for this bias (which, in the case of credit scoring, they refer to as reject inference). See, e.g., Jonathan Crook & John Banasik, *Does Reject Inference Really Improve the Performance of Application Scoring Models?*, 28 J. BANKING & FIN. 857 (2004).
C. Feature Selection

Through a process called “feature selection,” organizations—and the data miners that work for them—make choices about what attributes they observe and subsequently fold into their analyses. These decisions can also have serious implications for the treatment of protected classes if those factors that better account for pertinent statistical variation among members of a protected class are not well represented in the set of selected features. Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve.

This problem arises because data are necessarily reductive representations of an infinitely more specific real-world object or phenomenon. These representations may fail to capture enough detail to allow for the discovery of crucial points of contrast. Increasing the resolution and range of the analysis may still fail to capture the mechanisms that account for different outcomes because such mechanisms may not lend themselves to exhaustive or effective representation in the data, if such representations even exist. As Professors Toon Calders and Indrė Žliobaitė explain, “[I]t is often impossible to collect all the attributes of a subject or take all the environmental factors into account with a model.” While these limitations lend credence to the argument that a dataset can never fully encompass the full complexity of the individuals it seeks to represent, they do not reveal the inherent inadequacy of representation as such.

At issue, really, are the coarseness and comprehensiveness of the criteria that permit statistical discrimination and the uneven rates at which different groups happen to be subject to erroneous determinations. Crucially, these erroneous and potentially adverse outcomes are artifacts of statistical reasoning rather than prejudice on the part of decision makers or bias in the composition of the dataset. As Professor Frederick Schauer explains, decision makers that rely on statistically sound but nonuniversal generalizations “are being simultaneously rational and unfair” because certain individuals are “actuarially saddled” by statistically sound inferences that are nevertheless inaccurate.

64. Toon Calders & Indrė Žliobaitė, Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures, in Discrimination and Privacy in the Information Society, supra note 33, at 43, 46 (“[T]he selection of attributes by which people are described in [a] database may be incomplete.”).
66. Calders & Žliobaitė, supra note 64, at 47.
67. Frederick Schauer, Profiles, Probabilities, and Stereotypes 3–7 (2006). Insurance offers the most obvious example of this: the rate that a person pays for car insurance, for
Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs and may justify a less granular and encompassing analysis.68

To take an obvious example from the employment context, hiring decisions that consider academic credentials tend to assign enormous weight to the reputation of the college or university from which an applicant has graduated, even though such reputations may communicate very little about the applicant’s job-related skills and competencies.69 If equally competent members of protected classes happen to graduate from these colleges or universities at disproportionately low rates, decisions that turn on the credentials conferred by these schools, rather than some set of more specific qualities that more accurately sort individuals, will incorrectly and systematically discount these individuals. Even if employers have a rational incentive to look beyond credentials and focus on criteria that allow for more precise and more accurate determinations, they may continue to favor credentials because they communicate pertinent information at no cost to the employer.70

Similar dynamics seem to account for the practice known as “redlining,”71 in which financial institutions employ especially general criteria to draw distinctions between subpopulations (i.e., the neighborhood in which individuals happen to reside), despite the fact that such distinctions fail to capture significant variation within each subpopulation that would result in a different assessment for certain members of these groups. While redlining in America is well known to have had its basis in racial animus and prejudice,72 decision makers operating in this manner may attempt to justify their behavior by pointing to the cost efficiency of relying on easily accessible information. In other words, decision makers can argue that they are willing to tolerate higher rates of erroneous determinations for certain groups because the benefits

68. Kasper Lippert-Rasmussen, “We Are All Different”: Statistical Discrimination and the Right to Be Treated as an Individual, 15 J. ETHICS 47, 54 (2011) (“[O]btaining information is costly, so it is morally justified, all things considered, to treat people on the basis of statistical generalizations even though one knows that, in effect, this will mean that one will treat some people in ways, for better or worse, that they do not deserve to be treated.”); see also Brian Dalessandro, Claudia Perlich & Troy Raeder, Bigger Is Better, but at What Cost?: Estimating the Economic Value of Incremental Data Assets, 2 BIG DATA 87 (2014).


70. As one commentator has put it in contemplating data-driven hiring, “Big Data has its own bias. . . . You measure what you can measure.” Id.

71. See generally DAVID M. P. FREUND, COLORED PROPERTY: STATE POLICY AND WHITE RACIAL POLITICS IN SUBURBAN AMERICA (2010).

72. Id.
derived from more granular data—and thus better accuracy—do not justify the costs. Of course, it may be no coincidence that such cost-benefit analyses seem to justify treating groups composed disproportionately of members of protected classes to systematically less accurate determinations. Redlining is illegal because it can systematically discount entire areas composed primarily of members of a protected class, despite the presence of some qualified candidates.

Cases of so-called rational racism are really just a special instance of this more general phenomenon—one in which race happens to be taken into consideration explicitly. In such cases, decision makers take membership in a protected class into account, even if they hold no prejudicial views, because such membership seems to communicate relevant information that would be difficult or impossible to obtain otherwise. Accordingly, the persistence of distasteful forms of discrimination may be the result of a lack of information, rather than a continued taste for discrimination. Professor Lior Strahilevitz has argued, for instance, that when employers lack access to criminal records, they may consider race in assessing an applicant’s likelihood of having a criminal record because there are statistical differences in the rates at which members of different racial groups have been convicted of crimes. In other words, employers fall back on more immediately available and coarse features when they cannot access more specific or verified information. Of course, as Strahilevitz points out, race is a highly imperfect basis upon which to predict an individual’s criminal record, despite whatever differences may exist in the rates at which members of different racial groups have been convicted of crimes, because it is too coarse as an indicator.

73. While animus was likely the main motivating factor for redlining, the stated rationales were economic and about housing value. See Douglas S. Massey & Nancy A. Denton, American Apartheid: Segregation and the Making of the Underclass 51–52 (1993). Redlining persists today and may actually be motivated by profit, but it has the same deleterious effects. See Rachel L. Swarns, Biased Lending Evolves, and Blacks Face Trouble Getting Mortgages, N.Y. Times (Oct. 30 2015), http://www.nytimes.com/2015/10/31/nyregion/hudson-city-bank-settlement.html [https://perma.cc/P4YX-NTT9].

74. See Nationwide Mut. Ins. Co. v. Cisneros, 52 F.3d 1351, 1359 (6th Cir. 1995) (holding that the Fair Housing Act prohibited redlining in order “to eliminate the discriminatory business practices which might prevent a person economically able to do so from purchasing a house regardless of his race”); NAACP v. Am. Family Mut. Ins. Co., 978 F.2d 287, 300 (7th Cir. 1992).

75. See generally Andrea Romei & Salvatore Ruggieri, Discrimination Data Analysis: A Multi-Disciplinary Bibliography, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, supra note 33, at 109, 120.


77. Id. This argument assumes that criminal records are relevant to employment, which is often not true. See infra text accompanying note 175.

78. Strahilevitz, supra note 76, at 364; see also infra Part II.A. The law holds that decision makers should refrain from considering membership in a protected class even if statistical evidence seems to support certain inferences on that basis. The prohibition does not depend on whether decision
D. Proxies

Cases of decision making that do not artificially introduce discriminatory effects into the data mining process may nevertheless result in systematically less favorable determinations for members of protected classes. This is possible when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job—as formalized in some fashion—may also sort individuals according to class membership.

In certain cases, there may be an obvious reason for this. Just as “mining from historical data may . . . discover traditional prejudices that are endemic in reality (i.e., taste-based discrimination),” so, too, may data mining “discover patterns of lower performances, skills or capacities of protected-by-law groups.” These discoveries not only reveal the simple fact of inequality, but they also reveal that these are inequalities in which members of protected classes are frequently in the relatively less favorable position. This has rather obvious implications: if features held at a lower rate by members of protected groups nevertheless possess relevance in rendering legitimate decisions, such decisions will necessarily result in systematically less favorable determinations for these individuals. For example, by conferring greater attention and opportunities to employees that they predict will prove most competent at some task, employers may find that they subject members of protected groups to consistently disadvantageous treatment because the criteria that determine the attractiveness of employees happen to be held at systematically lower rates by members of these groups.

Decision makers do not necessarily intend this disparate impact because they hold prejudicial beliefs; rather, their reasonable priorities as profit seekers unintentionally recapitulate the inequality that happens to exist in society. Furthermore, this may occur even if proscribed criteria have been removed from the dataset, the data are free from latent prejudice or bias, the features are especially granular and diverse, and the only goal is to maximize classificatory or predictive accuracy. The problem stems from what researchers call “redundant encodings,” cases in which membership in a protected class happens to be encoded in other data. This occurs when a particular piece of data or certain values for that piece of data are highly correlated with

---

79. Romei & Ruggieri, supra note 75, at 121.
membership in specific protected classes. Data’s significant statistical relevance to the decision at hand helps explain why data mining can result in seemingly discriminatory models even when its only objective is to ensure the greatest possible accuracy for its determinations. If there is a disparate distribution of an attribute, a more precise form of data mining will be more likely to capture that distribution. Better data and more features will simply come closer to exposing the exact extent of inequality.

E. Masking

Data mining could also breathe new life into traditional forms of intentional discrimination because decision makers with prejudicial views can mask their intentions by exploiting each of the mechanisms enumerated above. Stated simply, any form of discrimination that happens unintentionally can also be orchestrated intentionally. For instance, decision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes.\(^{82}\) They could likewise attempt to preserve the known effects of prejudice in prior decision making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule. And decision makers could intentionally rely on features that only permit coarse-grained distinction making—distinctions that result in avoidably higher rates of erroneous determinations for members of a protected class. In denying themselves finer-grained detail, decision makers would be able to justify writing off entire groups composed disproportionately of members of protected classes. A form of digital redlining, this decision masks efforts to engage in intentional discrimination by abstracting to a level of analysis that fails to capture lower level variations. As a result, certain members of protected classes might not be seen as attractive candidates. Here, prejudice rather than some legitimate business reason (such as cost) motivates decision makers to intentionally restrict the particularity of their decision making to a level that can only paint in avoidably broad strokes. This condemns entire groups, composed disproportionately of members of protected classes, to systematically less favorable treatment.

Because data mining holds the potential to infer otherwise unseen attributes, including those traditionally deemed sensitive,\(^{83}\) it can indirectly determine individuals’ membership in protected classes and unduly discount, penalize, or exclude such people accordingly. In other words, data mining could grant decision makers the ability to distinguish and disadvantage members of protected classes even if those decision makers do not have access to explicit information about individuals’ class membership. Data mining could

\(^{82}\) See id. (discussing the “[s]elf-fulfilling prophecy”).

instead help to pinpoint reliable proxies for such membership and thus place institutions in the position to automatically sort individuals into their respective class without ever having to learn these facts directly. The most immediate implication is that institutions could employ data mining to circumvent the barriers, both practical and legal, that have helped to withhold individuals’ protected class membership from consideration.

Additionally, data mining could provide cover for intentional discrimination of this sort because the process conceals the fact that decision makers determined and considered the individual’s class membership. The worry, then, is not simply that data mining introduces novel ways for decision makers to satisfy their taste for illegal discrimination; rather, the worry is that it may mask actual cases of such discrimination. Although scholars, policy makers, and lawyers have long been aware of the dangers of masking, data mining significantly enhances the ability to conceal acts of intentional discrimination by finding ever more remote and complex proxies for proscribed criteria.

Intentional discrimination and its masking have so far garnered disproportionate attention in discussions of data mining, often to the exclusion of issues arising from the many forms of unintentional discrimination described above. While data mining certainly introduces novel ways to discriminate intentionally and to conceal those intentions, most cases of employment discrimination are already sufficiently difficult to prove; employers motivated by conscious prejudice would have little to gain by pursuing these complex and costly mechanisms to further mask their intentions. When it comes to data mining, unintentional discrimination is the more pressing concern because it is likely to be far more common and easier to overlook.

84. Id. at 9–13.
85. Data miners who wish to discriminate can do so using relevant or irrelevant criteria. Either way the intent would make the action “masking.” If an employer masked using highly relevant data, litigation arising from it likely would be tried under a “mixed-motive” framework, which asks whether the same action would have been taken without the intent to discriminate. See infra Part II.A.
86. See, e.g., Custers, supra note 33, at 9–10.
87. See Barocas, supra note 83.
88. See, e.g., Alistair Croll, Big Data Is Our Generation’s Civil Rights Issue, and We Don’t Know It, SOLVE FOR INTERESTING (July 31, 2012, 12:40 PM), http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it [https://perma.cc/BS8S-6T7S]. This post generated significant online chatter immediately upon publication and has become one of the canonical texts in the current debate. It has also prompted a number of responses from scholars. See, e.g., Anders Sandberg, Asking the Right Questions: Big Data and Civil Rights, PRAC. ETHICS (Aug. 16, 2012), http://blog.practicalethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights [https://perma.cc/NC36-NBZN].
II.

TITLE VII LIABILITY FOR DISCRIMINATORY DATA MINING

Current antidiscrimination law is not well equipped to address the cases of discrimination stemming from the problems described in Part I. This Part considers how Title VII might apply to these cases. Other antidiscrimination laws, such as the Americans with Disabilities Act, will exhibit differences in specific operation, but the main thrust of antidiscrimination law is fairly consistent across regimes, and Title VII serves as an illustrative example.90

An employer sued under Title VII may be found liable for employment discrimination under one of two theories of liability: disparate treatment and disparate impact.91 Disparate treatment comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate.92 Disparate impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes.93 Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.94

Liability under Title VII for discriminatory data mining will depend on the particular mechanism by which the inequitable outcomes are generated. This Part explores the disparate treatment and disparate impact doctrines and analyzes which mechanisms could generate liability under each theory.

A. Disparate Treatment

Disparate treatment recognizes liability for both explicit formal classification and intentional discrimination.95 Formal discrimination, in which membership in a protected class is used as an input to the model, corresponds to an employer classifying employees or potential hires according to membership in a protected class and differentiating them on that basis. Formal

90. The biggest difference between the Americans with Disabilities Act and Title VII is the requirement that an employer make “reasonable accommodations” for disabilities. 42 U.S.C. § 12112(b)(5) (2012). But some scholars have argued that even this difference is illusory and that accommodations law is functionally similar to Title VII, though worded differently. See Samuel R. Bagenstos, “Rational Discrimination,” Accommodation, and the Politics of (Disability) Civil Rights, 89 Va. L. Rev. 825, 833 & n.15 (2003) (comparing accommodations law to disparate treatment); Christine Jolls, Antidiscrimination and Accommodation, 115 Harv. L. Rev. 642, 652 (2001) (comparing accommodations law to disparate impact).


95. Id. § 2000e-2(a), (k); see Primus, supra note 92, at 1350–51 n.56.
discrimination covers both the straightforward denial of opportunities based on protected class membership and the use of rational racism. In traditional contexts, rational racism is considered rational because there are cases in which its users believe it is an accurate, if coarse-grained, proxy—or at least the best available one in a given situation. In the world of data mining, though, that need not be the case. Even if membership in a protected class were specified as an input, the eventual model that emerges could see it as the least significant feature. In that case, there would be no discriminatory effect, but there would be a disparate treatment violation, because considering membership in a protected class as a potential proxy is a legal classificatory harm in itself.

Formal liability does not correspond to any particular discrimination mechanism within data mining; it can occur equally well in any of them. Because classification itself can be a legal harm, irrespective of the effect, the same should be true of using protected class as an input to a system for which the entire purpose is to build a classificatory model. The irony is that the use of protected class as an input is usually irrelevant to the outcome in terms of discriminatory effect, at least given a large enough number of input features. The target variable will, in reality, be correlated to the membership in a protected class somewhere between 0 percent and 100 percent. If the trait is perfectly uncorrelated, including membership in the protected class as an input will not change the output, and there will be no discriminatory effect. On the other end of the spectrum, where membership in the protected class is perfectly predictive of the target variable, the fact will be redundantly encoded in the other data. The only way using membership in the protected class as an explicit feature will change the outcome is if the information is otherwise not rich enough to detect such membership. Membership in the protected class will prove relevant to the exact extent it is already redundantly encoded. Given a rich enough set of features, the chance that such membership is redundantly encoded approaches certainty. Thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input. Formal discrimination therefore should have no bearing whatsoever on the

100. Membership in a protected class is still a permissible input to a holistic determination when the focus is diversity, but where classification is the goal, such as here, it is not. See Grutter v. Bollinger, 539 U.S. 306, 325 (2003) (noting that “diversity is a compelling state interest” that can survive strict scrutiny).
101. That is, not counting any expressive harm that might come from classification by protected class.
outcome of the model. Additionally, by analyzing the data, an employer could probabilistically determine an employee’s membership in that same protected class, if the employer did indeed want to know.

To analyze intentional discrimination other than mere formal discrimination, a brief description of disparate treatment doctrine is necessary. A Title VII disparate treatment case will generally proceed under either the McDonnell-Douglas burden-shifting scheme or the Price-Waterhouse “mixed motive” regime. Under the McDonnell-Douglas framework, the plaintiff who has suffered an adverse employment action has the initial responsibility to establish a prima facie case of discrimination by demonstrating that a similarly situated person who is not a member of a protected class would not have suffered the same fate. This can be shown with circumstantial evidence of discriminatory intent, such as disparaging remarks made by the employer or procedural irregularities in promotion or hiring; only very rarely will an employer openly admit to discriminatory conduct. If the plaintiff successfully demonstrates that the adverse action treated protected class members differently, then the burden shifts to the defendant-employer to offer a legitimate, nondiscriminatory basis for the decision. The defendant need not prove the reason is true; his is only a burden of production. Once the defendant has offered a nondiscriminatory alternative, the ultimate burden of persuasion falls to the plaintiff to demonstrate that the proffered reason is pretextual.

In the data mining context, liability for masking is clear as a theoretical matter, no matter which mechanism for discrimination is employed. The fact that it is accomplished algorithmically does not make it less of a disparate treatment violation, as the entire idea of masking is pretextual. In fact, in the traditional, non–data mining context, the word masking has occasionally been used to refer to pretext. Like in any disparate treatment case, however, proof will be difficult to come by, something even truer for masking.


103. This is similar to the computer science definition of discrimination. Calders & Žliobaite, supra note 64, at 49. (“A classifier discriminates with respect to a sensitive attribute, e.g. gender, if for two persons which only differ by their gender (and maybe some characteristics irrelevant for the classification problem at hand) that classifier predicts different labels.”).


105. Id.

106. See Keyes v. Sec’y of the Navy, 853 F.2d 1016, 1026 (1st Cir. 1988) (explaining that it is the plaintiff’s burden to show that the proffered reasons for hiring an alternative were “pretexts aimed at masking sex or race discrimination”); Custers, supra note 33, at 9–10; Megan Whitehill, Better Safe than Subjective: The Problematic Intersection of Prehire Social Networking Checks and Title VII Employment Discrimination, 85 TEMP. L. REV. 229, 250 (2012) (referring to “[m]asking [p]retext” in the third stage of McDonnell-Douglas framework).

107. See supra Part I.E. This is a familiar problem to antidiscrimination law, and it is often cited as one of the rationales for disparate impact liability in the first place—to “smoke out” intentional invidious discrimination. See infra Part III.B.
The McDonnell-Douglas framework operates on a presumption that if the rationale that the employer has given is found to be untrue, the employer must be holding his “true” discriminatory motive.\(^\text{108}\) Because the focus of the McDonnell-Douglas framework is on pretext and cover-up, it can only address conscious, willful discrimination.\(^\text{109}\) Under the McDonnell-Douglas framework, a court must find either that the employer intended to discriminate or did not discriminate at all.\(^\text{110}\) Thus, unintentional discrimination will not lead to liability.

A Title VII disparate treatment case can also be tried under the mixed-motive framework, first recognized in Price Waterhouse v. Hopkins\(^\text{111}\) and most recently modified by Desert Palace, Inc. v. Costa.\(^\text{112}\) In the mixed-motive framework, a plaintiff need not demonstrate that the employer’s nondiscriminatory rationale was pretextual, but merely that discrimination was a “motivating factor” in the adverse employment action.\(^\text{113}\) As a practical matter, this means that the plaintiff must show that the same action would not have been taken absent the discriminatory motive.\(^\text{114}\) As several commentators

\begin{itemize}
  \item \(^{108}\) McDonnell Douglas Corp. v. Green, 411 U.S. 792, 805 (1973) (The plaintiff “must be given a full and fair opportunity to demonstrate by competent evidence that the presumptively valid reasons for his rejection were in fact a cover-up for a racially discriminatory decision”). While, as a theoretical matter, the plaintiff must prove that the employer’s reason was a pretext for discrimination specifically, the Supreme Court has held that a jury can reasonably find that the fact that an employer had only a pretextual reason to fall back on is itself circumstantial evidence of discrimination. Hicks, 509 U.S. at 511 (“The factfinder’s disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.”).
  \item \(^{109}\) See Tristin K. Green, Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory, 38 HARV. C.R.-C.L. L. REV. 91, 114 (2003) (“Presuming that individuals know the real reason for their actions, the pretext model of disparate treatment provides that an employer can be held to have discriminated when the plaintiff establishes a minimal prima facie case and shows that the reason given for the adverse decision is unworthy of credence.”); Susan Sturm, Second Generation Employment Discrimination: A Structural Approach, 101 COLUM. L. REV. 458, 458 (2001); see also Melissa Hart, Subjective Decisionmaking and Unconscious Discrimination, 56 ALA. L. REV. 741, 749–50 (2005) (critiquing the courts’ requirement of proving employer “dishonesty,” but suggesting that, absent this requirement, Title VII could handle unconscious discrimination without altering the law).
  \item \(^{110}\) Krieger, supra note 89, at 1170.
  \item \(^{111}\) 490 U.S. 228 (1989).
  \item \(^{112}\) 539 U.S. 90 (2003).
  \item \(^{113}\) 42 U.S.C. § 2000e-2(m) (2012); Desert Palace, 539 U.S. at 101 (“In order to obtain [a mixed-motive jury instruction], a plaintiff need only present sufficient evidence for a reasonable jury to conclude, by a preponderance of the evidence, that ‘race, color, religion, sex, or national origin was a motivating factor for any employment practice.’”). The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.
\end{itemize}
have pointed out, motive and intent are not necessarily synonymous. Motive can be read more broadly to include unconscious discrimination, including anything that influences a person to act, such as emotions or desires. Nonetheless, courts have conflated the meanings of motive and intent such that the phrase “motive or intent” has come to refer only to conscious choices. Thus, while most individual decision making probably belongs in a mixed-motive framework, as each decision a person makes comprises a complicated mix of motivations, the mixed-motive framework will be no better than the pretext framework at addressing bias that occurs absent conscious intent.

Except for masking, discriminatory data mining is by stipulation unintentional. Unintentional disparate treatment is not a problem that is new to data mining. A vast scholarly literature has developed regarding the law’s treatment of unconscious, implicit bias. Such treatment can occur when an employer has internalized some racial stereotype and applies it or, without realizing it, monitors an employee more closely until the employer finds a violation. The employee is clearly treated differently, but it is not intentional, and the employer is unaware of it. As Professor Samuel Bagenstos summarized, at this point, “it may be difficult, if not impossible, for a court to go back and reconstruct the numerous biased evaluations and perceptions that ultimately resulted in an adverse employment decision.” Within the scholarly literature, there is “[s]urprising unanimity” that the law does not adequately address unconscious disparate treatment.

115. Krieger, supra note 89, at 1243; Sullivan, supra note 114, at 915.
116. Krieger, supra note 89, at 1243; Sullivan, supra note 114, at 915 n.18 (quoting Motive, OXFORD ENGLISH DICTIONARY (1st ed. 1933)).
117. Sullivan, supra note 114, at 914–16, 916 n.20.
119. See Krieger, supra note 89, at 1182–83.
121. This example can be ported directly to data mining as overrepresentation in data collection. See supra Part I.B.2.
123. Sullivan, supra note 114, at 1000. There is, however, no general agreement on whether the law should treat such discrimination as disparate treatment or disparate impact. Compare Krieger, supra note 89, at 1231 (explaining that because the bias causes employers to treat people differently, it
There are a few possible ways to analogize discriminatory data mining to unintentional disparate treatment in the traditional context, based on where one believes the “treatment” lies. Either the disparate treatment occurs at the decision to apply a predictive model that will treat members of a protected class differently, or it occurs when the disparate result of the model is used in the ultimate hiring decision. In the first scenario, the intent at issue is the decision to apply a predictive model with known disproportionate impact on protected classes. In the second, the disparate treatment occurs if, after the employer sees the disparate result, he proceeds anyway. If the employer continues because he liked the discrimination produced in either scenario, then intent is clear. If not, then this just devolves into a standard disparate impact scenario, with liability based on effect. Under disparate impact theory, deciding to follow through on a test with discriminatory effect does not suddenly render it disparate treatment.124

Another option is to imagine the model as the decision maker exhibiting implicit bias. That is, because of biases hidden to the predictive model such as nonrepresentative data or mislabeled examples, the model reaches a discriminatory result. This analogy turns every mechanism except proxy discrimination into the equivalent of implicit bias exhibited by individual decision makers. The effect of bias is one factor among the many different factors that go into the model-driven decision, just like in an individual’s adverse employment decision.125 Would a more expansive definition of motive fix this scenario?

Because the doctrine focuses on human decision makers as discriminators, the answer is no. Even if disparate treatment doctrine could capture unintentional discrimination, it would only address such discrimination stemming from human bias. For example, the person who came up with the idea for Street Bump ultimately devised a system that suffers from reporting bias,126 but it was not because he or she was implicitly employing some racial stereotype. Rather, it was simply inattentiveness to problems with the sampling frame. This is not to say that his or her own bias had nothing to do with it—the person likely owned a smartphone and thus did not think about the people who do not—but no one would say that it was even implicit bias against protected

---

124. In fact, after Ricci v. DeStefano, 557 U.S. 557 (2009), deciding not to apply such a test after noticing the discriminatory effect may give rise to a disparate treatment claim in the other direction.

125. Bagenstos, supra note 122, at 9; Krieger, supra note 89, at 1185–86 (“Not only disparate treatment analysis, but the entire normative structure of Title VII’s injunction ‘not to discriminate,’ rests on the assumption that decisionmakers possess ‘transparency of mind’—that they are aware of the reasons why they are about to make, or have made, a particular employment decision.”).

126. See supra note 51 and accompanying text.
classes that motivated the decision, even under the expansive definition of the word “motive.” 127

The only possible analogy relevant to disparate treatment, then, is to those data mining mechanisms of unintentional discrimination that reflect a real person’s bias—something like LinkedIn’s Talent Match recommendation engine, which relies on potentially prejudiced human assessments of employees. 128 As a general rule, an employer may not avoid disparate treatment liability by encoding third-party preferences as a rationale for a hiring decision. 129 But, once again, to be found liable under current doctrine, the employer would likely both have to know that this is the specific failure mechanism of the model and choose it based on this fact.

There is one other interesting question regarding disparate treatment doctrine: whether the intent standard includes knowledge. This is not a problem that arises often when a human is making a single employment determination. Assuming disparate treatment occurs in a given case, it is generally either intended or unconscious. What would it mean to have an employer know that he was treating an employee differently, but still take the action he had always planned to take without intent to treat the employee differently? It seems like an impossible line to draw. 130

With data mining, though, unlike unconscious bias, it is possible to audit the resulting model and inform an employer that she will be treating individuals differently before she does so. If an employer intends to employ the model, but knows it will produce a disparate impact, does she intend to discriminate? This is a more realistic parsing of intent and knowledge than in the case of an individual, nonsystematic employment decision. Neither pretext nor motive exists here, and throughout civil and criminal law, “knowledge” and “intent” are considered distinct states of mind, so there would likely be no liability. On the other hand, courts may use knowledge of discrimination as evidence to find intent. 131 And while the statute’s language only covers intentional discrimination, a broad definition of intent could include knowledge or

127. Of course, the very presumption of a design’s neutrality is itself a bias that may work against certain people. See Langdon Winner, Do Artifacts Have Politics?, 109 DAEDALUS 121, 125 (1980). But, as this is a second-order effect, we need not address it here.

128. See Woods, supra note 46.

129. See 29 C.F.R. § 1604.2(a)(1)(ii) (2015) (stating the EEOC’s position that “the preferences of coworkers, the employer, clients or customers” cannot be used to justify disparate treatment); see also Fernandez v. Wynn Oil Co., 653 F.2d 1273, 1276–77 (9th Cir. 1981); Diaz v. Pan Am. World Airways, Inc., 442 F.2d 385, 389 (5th Cir. 1971).

130. See Krieger, supra note 89, at 1185 (discussing disparate treatment’s “assumption of decisionmaker self-awareness”).

131. Columbus Bd. of Educ. v. Penick, 443 U.S. 449, 464 (1979) (“[A]ctions having foreseeable and anticipated disparate impact are relevant evidence to prove the ultimate fact, forbidden purpose.”); Pers. Adm’r of Mass. v. Feeney, 442 U.S. 256, 279 n.25 (1979) (“[W]hen the adverse consequences of a law upon an identifiable group are . . . inevitable . . . , a strong inference that the adverse effects were desired can reasonably be drawn.”).

substantial certainty of the result. Because the situation has not come up often, the extent of the “intent” required is as yet unknown.

In sum, aside from rational racism and masking (with some difficulties), disparate treatment doctrine does not appear to do much to regulate discriminatory data mining.

B. Disparate Impact

Where there is no discriminatory intent, disparate impact doctrine should be better suited to finding liability for discrimination in data mining. In a disparate impact case, a plaintiff must show that a particular facially neutral employment practice causes a disparate impact with respect to a protected class. If shown, the defendant-employer may “demonstrate that the challenged practice is job related for the position in question and consistent with business necessity.” If the defendant makes a successful showing to that effect, the plaintiff may still win by showing that the employer could have used an “alternative employment practice” with less discriminatory results.

The statute is unclear as to the required showing for essentially every single element of a disparate impact claim. First, it is unclear how much disparate impact is needed to make out a prima facie case. The EEOC, charged with enforcing Title VII’s mandate, has created the so-called “four-fifths rule” as a presumption of adverse impact: “A selection rate for any race, sex, or ethnic group which is less than four-fifths . . . of the rate for the group.

---

133. See Julia Kobick, Note, Discriminatory Intent Reconsidered: Folk Concepts of Intentionality and Equal Protection Jurisprudence, 45 HARV. C.R.-C.L. L. REV. 517, 551 (2010) (arguing that courts should regularly consider knowledge and foreseeability of disparate impact as an intended effect); cf. RESTATEMENT (SECOND) OF TORTS § 8A cmt. b (AM. LAW INST. 1965) (“Intent is not . . . limited to consequences which are desired. If the actor knows that the consequences are certain, or substantially certain, to result from his act, and still goes ahead, he is treated by the law as if he had in fact desired to produce the result.”).

134. Determining that a model is discriminatory is also like trying and failing to validate a test under disparate impact doctrine. See infra Part II.B. If a test fails validation, the employer using it would know that he is discriminating if he applies it, but that does not imply that he is subject to disparate treatment liability. Nonetheless, validation is part of the business necessity defense, and that defense is not available against disparate treatment claims. Thus, the analysis does not necessarily have the same result. 42 U.S.C. § 2000e-2(k)(2). One commentator has argued that including knowledge as a state of mind leading to disparate treatment liability would effectively collapse disparate impact and disparate treatment by conflating intent and effect. Jessie Allen, A Possible Remedy for Unthinking Discrimination, 61 BROOK. L. REV. 1299, 1314 (1995). But others still have noted that with respect to knowledge, a claim is still about the treatment of an individual, not the incidental disparate impact of a neutral policy. See Carin Ann Claus, Comparable Worth—The Theory, Its Legal Foundation, and the Feasibility of Implementation, 20 U. MICH. J.L. REFORM 7, 62 (1986).


136. Id.

137. Id.

138. The statute does not define the requirement and Supreme Court has never addressed the issue. See, e.g., Sullivan, supra note 114, at 954 & n.153. For a brief discussion of the different approaches to establishing disparate impact, see Pamela L. Perry, Two Faces of Disparate Impact Discrimination, 59 FORDHAM L. REV. 523, 570–74 (1991).
with the highest rate will generally be regarded . . . as evidence of adverse impact.”\textsuperscript{139} The Uniform Guidelines on Employment Selection Procedures (Guidelines) also state, however, that smaller differences can constitute adverse impact and greater differences may not, depending on circumstances. Thus, the four-fifths rule is truly just a guideline.\textsuperscript{140} For the purposes of this Part, it is worthwhile to just assume that the discriminatory effects are prominent enough to establish disparate impact as an initial matter.\textsuperscript{141}

The next step in the litigation is the “business necessity” defense. This defense is, in a very real sense, the crux of disparate impact analysis, weighing Title VII’s competing goals of limiting the effects of discrimination while allowing employers discretion to advance important business goals. Griggs v. Duke Power Co.\textsuperscript{142}—the decision establishing the business necessity defense alongside disparate impact doctrine itself—articulated the defense in several different ways:

A challenged employment practice must be “shown to be related to job performance,” have a “manifest relationship to the employment in question,” be “demonstrably a reasonable measure of job performance,” bear some “relationship to job-performance ability,” and/or “must measure the person for the job and not the person in the abstract.”\textsuperscript{143}

The Supreme Court was not clear on what, if any, difference existed between job-relatedness and business necessity, at one point seeming to use the terms interchangeably: “The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited.”\textsuperscript{144} The focus of the Court was clearly on future job performance, and the term “job-related” has come to mean a practice that is predictive of job performance.\textsuperscript{145} Because the definitions of job-relatedness and business necessity have never been clear, courts defer when applying the doctrine and finding the appropriate balance.\textsuperscript{146}

Originally, the business necessity defense seemed to apply narrowly. In Griggs, Duke Power had instituted new hiring requirements including a high school diploma and success on a “general intelligence” test for previously

\textsuperscript{140} Id.
\textsuperscript{141} We will return to this when discussing the need to grapple with substantive fairness. See infra Part III.B.
\textsuperscript{142} 401 U.S. 424 (1971).
\textsuperscript{144} Griggs, 401 U.S. at 431; see also Lye, supra note 143, at 320.
\textsuperscript{145} Lye, supra note 143, at 355 & n.206.
white-only divisions. Duke Power did not institute such requirements in divisions where it had previously hired black employees.\textsuperscript{147} The Court ruled that the new requirements were not a business necessity because “employees who have not completed high school or taken the tests have continued to perform satisfactorily and make progress in departments for which the high school and test criteria are now used.”\textsuperscript{148} Furthermore, the requirements were implemented without any study of their future effect.\textsuperscript{149} The Court also rejected the argument that the requirements would improve the “overall quality of the workforce.”\textsuperscript{150}

By 1979, the Court began treating business necessity as a much looser standard.\textsuperscript{151} In \textit{New York City Transit Authority v. Beazer},\textsuperscript{152} the transit authority had implemented a rule barring drug users from employment, including current users of methadone, otherwise known as \textit{recovering} heroin addicts. In dicta, the Court stated that a “narcotics rule,” which “significantly serves” the “legitimate employment goals of safety and efficiency,” was “assuredly” job related.\textsuperscript{153} This was the entire analysis of the business necessity defense in the case. Moreover, the rationale was acceptable as applied to the entire transit authority, even where only 25 percent of the jobs were labeled as “safety sensitive.”\textsuperscript{154} Ten years later, the Court made the business necessity doctrine even more defendant-friendly in \textit{Wards Cove Packing Co. v. Atonio}.\textsuperscript{155}

After \textit{Wards Cove}, the business necessity defense required a court to engage in “a reasoned review of the employer’s justification for his use of the challenged practice. . . . [T]here is no requirement that the challenged practice be ‘essential’ or ‘indispensable’ to the employer’s business for it to pass muster . . . .”\textsuperscript{156} The Court also reallocated the burden to plaintiffs to prove that business necessity was lacking and even referred to the defense as a “business justification” rather than a business necessity.\textsuperscript{157} The \textit{Wards Cove} Court went so far that Congress directly addressed the decision in the Civil Rights Act of 1991 (1991 Act), which codified disparate impact and reset the standards to the day before \textit{Wards Cove} was decided.\textsuperscript{158}

Because the substantive standards for job-relatedness or business necessity were uncertain before \textit{Wards Cove}, however, the confusion persisted

\begin{itemize}
\item 147. Griggs, 401 U.S. at 427–28.
\item 148. Id. at 431–32.
\item 149. Id. at 432.
\item 150. Id. at 431.
\item 152. 440 U.S. 568 (1979).
\item 153. Id. at 587 & n.31.
\item 154. Id.
\item 155. 490 U.S. 642 (1989).
\item 156. Id. at 659.
\item 157. Id.
\end{itemize}
even after the 1991 Act was passed. At the time, both sides—civil rights groups and the Bush administration, proponents of a rigorous and more lenient business necessity defense respectively—declared victory.

Since then, courts have recognized that business necessity lies somewhere in the middle of two extremes. Some courts require that the hiring criteria bear a “manifest relationship” to the employment in question or that they be “significantly correlated” to job performance. The Third Circuit was briefly an outlier, holding “that hiring criteria must effectively measure the ‘minimum qualifications for successful performance of the job’” in order to meet the strict business necessity standard. This tougher standard would, as a practical matter, ban general aptitude tests with any disparate impact because a particular cutoff score cannot be shown to distinguish between those able and completely unable to do the work. For example, other unmeasured skills and abilities could theoretically compensate for the lower score on an aptitude test, rendering a certain minimum score not “necessary” if it does not measure minimum qualifications. In a subsequent case, however, the Third Circuit recognized that Title VII does not require an employer to choose someone “less qualified” (as opposed to unqualified) in the name of nondiscrimination and noted that aptitude tests can be legitimate hiring tools if they accurately measure a person’s qualifications. The court concluded:

159. Legislative history was no help either. The sole piece of legislative history is an “interpretive memorandum” that specifies that the standards were to revert to before Wards Cove, coupled with an explicit instruction in the Act to ignore any other legislative history regarding business necessity. Susan S. Grover, The Business Necessity Defense in Disparate Impact Discrimination Cases, 30 GA. L. REV. 387, 392–93 (1996).


161. Though courts generally state the standard to reflect this middle position, the Supreme Court’s latest word on disparate impact—in which the Court reaffirmed the doctrine generally and held that it applied in the Fair Housing Act—including the decidedly defendant-friendly observation that “private policies are not contrary to the disparate-impact requirement unless they are ‘artificial, arbitrary, and unnecessary barriers.’” Tex. Dep’t of Hous. & Cnty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507, 2512 (2015) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971)).

162. See, e.g., Gallagher v. Magna, 619 F.3d 823, 834 (8th Cir. 2010); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 265 (4th Cir. 2005).

163. Gulino v. N.Y. State Educ. Dep’t, 460 F.3d 361, 383 (2d Cir. 2006) (noting that hiring criteria are “significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated” (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 431 (1975))).


166. Id. Note, though, that this is similar to arguing that there is a less discriminatory alternative employment practice. This argument, then, would place the burden of the alternative employment practice prong on the defendant, contravening the burden-shifting scheme in the statute. See infra notes 170–74 and accompanying text.

Putting these standards together, then, we require that employers show that a discriminatory hiring policy accurately—but not perfectly—ascertains an applicant’s ability to perform successfully the job in question. In addition, Title VII allows the employer to hire the applicant most likely to perform the job successfully over others less likely to do so.\footnote{168} Thus, all circuits seem to accept varying levels of job-relatedness rather than strict business necessity.\footnote{169}

The last piece of the disparate impact test is the “alternative employment practice” prong. Shortly after Griggs, the Supreme Court decided Albemarle Paper Co. v. Moody, holding in part that “[i]f an employer does then meet the burden of proving that its tests are ‘job related,’ it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer’s legitimate interest in ‘efficient and trustworthy workmanship.’”\footnote{170} This burden-shifting scheme was codified in the 1991 Act as the “alternative employment practice” requirement.\footnote{171} Congress did not define the phrase, and its substantive meaning

\footnote{168} Id.
\footnote{169} Interestingly, it seems that many courts read identical business necessity language in the Americans with Disabilities Act to refer to a minimum qualification standard. See, e.g., Sullivan v. River Valley Sch. Dist., 197 F.3d 804, 811 (6th Cir. 1999) (“[T]here must be significant evidence that could cause a reasonable person to inquire as to whether an employee is still capable of performing his job. An employee’s behavior cannot be merely annoying or inefficient to justify an examination; rather, there must be genuine reason to doubt whether that employee can ‘perform job-related functions.’” (quoting 42 U.S.C. § 12112(d)(4)(B))). Presumably, this is because disability, when compared to race or sex, more immediately raises questions regarding a person’s ability to perform a job. Ironically, however, this means that disparate impact will be more tolerated where it is less likely to be obviously justified. Christine Jolls has in fact argued that disparate impact is, to a degree, functionally equivalent to accommodations law. Jolls, supra note 90, at 652.
\footnote{170} 42 U.S. 405, 425 (1975) (quoting McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973)).
\footnote{171} 42 U.S.C. § 2000e-2(k)(1)(A) (2012). The “alternative employment practice” test has not always been treated as a separate step. See, e.g., Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659 (1989) (treating the alternative employment practice test as part of the “business justification” phase); Dothard v. Rawlinson, 433 U.S. 321, 332 (1977) (treating the alternative employment practice test as a narrow tailoring requirement for the business necessity defense). The Albemarle Court, though creating a surrebuttal and thus empowering plaintiffs, seemed to regard the purpose of disparate impact as merely smoking out pretexts for intentional discrimination. 422 U.S. at 425; see also Primus, supra note 98, at 537. If the Albemarle Court’s approach is correct, treating the alternative employment practice requirement as a narrow tailoring requirement does make sense, much as the narrow tailoring requirement of strict scrutiny in equal protection serves the function of smoking out invidious purpose. City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493 (1989); Rubenfeld, supra note 99, at 428.

Every circuit to address the question, though, has held that the 1991 Act returned the doctrine to the Albemarle burden-shifting scheme. Jones v. City of Boston, 752 F.3d 38, 54 (1st Cir. 2014); Howe v. City of Akron, 723 F.3d 651, 658 (6th Cir. 2013); Tabor v. Hilti, Inc., 703 F.3d 1206, 1220 (10th Cir. 2013); Puffer v. Allstate Ins. Co., 675 F.3d 709, 717 (7th Cir. 2012); Gallagher v. Magnier, 619 F.3d 823, 833 (8th Cir. 2010); Gulino v. N.Y. State Educ. Dep’t, 460 F.3d 361, 382 (2d Cir. 2006); InT’l Bhd. of Elec. Workers Local Unions Nos. 605 & 985 v. Miss. Power & Light Co., 442 F.3d 313, 318 (5th Cir. 2006); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 277
remains uncertain. *Wards Cove* was the first case to use the specific phrase, so Congress’s instruction to reset the law to the pre-*Wards Cove* standard is particularly perplexing. The best interpretation is most likely *Albemarle’s* reference to “other tests or selection devices, without a similarly undesirable racial effect.” But this interpretation is slightly odd because in *Albemarle*, business necessity was still somewhat strict, and it is hard to imagine a business practice that is “necessary” while there exists a less discriminatory alternative that is just as effective. If business necessity or job-relatedness is a less stringent requirement, though, then the presence of the alternative employment practice requirement does at least give it some teeth.

Now return to data mining. For now, assume a court does not apply the strict business necessity standard but has some variation of “job related” in mind (as all federal appellate courts do today). The threshold issue is clearly whether the sought-after trait—the target variable—is job related, regardless of the machinery used to predict it. If the target variable is not sufficiently job related, a business necessity defense would fail, regardless of the fact that the decision was made by algorithm. Thus, disparate impact liability can be found for improper care in target variable definition. For example, it would be difficult for an employer to justify an adverse determination based on the appearance of an advertisement suggesting a criminal record alongside the search results for a candidate’s name. Sweeney found such a search to have a disparate impact, and the EEOC and several federal courts have interpreted Title VII to prohibit discrimination on the sole basis of criminal record, unless there is a specific reason the particular conviction is related to the job.

---


173. *Albemarle*, 422 U.S. at 425; accord, e.g., *Jones*, 752 F.3d at 53 (citing *Albemarle* to find meaning in the 1991 Act’s text); *Allen v. City of Chicago*, 351 F.3d 306, 312 (7th Cir. 2003) (same, but with a “see also” signal).


175. The difference would be whether mining for a single job-related trait, rather than a holistic ranking of “good employees,” is permissible at all. See infra text accompanying notes 197–99.

176. Sweeney, supra note 41, at 51.

is true independent of the fact that the disparity is an artifact of third-party bias; all that matters is whether the target variable is job related. In the end, though, because determining that a business practice is not job related actually requires a normative determination that it is instead discriminatory, courts tend to accept most common business practices for which an employer has a plausible story.178

Once a target variable is established as job related, the first question is whether the model is predictive of that trait. The nature of data mining suggests that this will be the case. Data mining is designed entirely to predict future outcomes, and, if seeking a job-related trait, future job performance. One commentator lamented that “[f]ederal case law has shifted from a prospective view of meritocracy to a retrospective view, thereby weakening disparate impact law.”179 The author meant that, in Griggs, the Court recognized that education and other external factors were unequal and therefore discounted a measure of meritocracy that looked to past achievements, in favor of comparing the likelihood of future ones. But by the time the Court had decided Wards Cove, it had shifted to a model of retrospective meritocracy that presumed the legitimacy of past credentials, thus upholding the status quo.180

While data mining must take the past—represented by the training data—as given, it generates predictions about workplace success that are much more accurate than predictions based on those past credentials that disparate impact doctrine has come to accept.181 In a hypothetical perfect case of data mining, the available information would be rich enough that reliance on the past information would fully predict future performance. Thus, robust data mining would likely satisfy even the Griggs Court’s standard that the models are looking toward future job performance, not merely past credentials.

The second question asks whether the model adequately predicts what it is supposed to predict. In the traditional context, this question arises in the case of general aptitude tests that might end up measuring unrelated elements of cultural awareness rather than intelligence.182 This is where the different data

---

179. DeSario, supra note 151, at 481.
180. Id. at 493; see also infra Conclusion.
182. See, e.g., Griggs v. Duke Power Co., 420 F.2d 1225, 1239 n.6 (4th Cir. 1970), rev’d, 401 U.S. 424 (1971) (“Since for generations blacks have been afforded inadequate educational opportunities and have been culturally segregated from white society, it is no more surprising that their
mining mechanisms for discriminatory effects matter. Part I posited that proxy discrimination optimizes correctly. So if it evidences a disparate impact, it reflects unequal distribution of relevant traits in the real world. Therefore, proxy discrimination will be as good a job predictor as possible given the current shape of society. Models trained on biased samples and mislabeled examples, on the other hand, will result in correspondingly skewed assessments rather than reflect real-world disparities. The same effect may be present in models that rely on insufficiently rich or insufficiently granular datasets: by designation they do not reflect reality. These models might or might not be considered job related, depending on whether the errors distort the outcomes enough that the models are no longer good predictors of job performance.

The Guidelines have set forth validation procedures intended to create a job-relatedness standard. Quantifiable tests that have a disparate impact must be validated according to the procedures in the Guidelines if possible; otherwise, a presumption arises that they are not job related. Under the Guidelines, a showing of validity takes one of three forms: criterion-related, content, or construct. Criterion-related validity “consist[s] of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance.” The “relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance.” Content validity refers to testing skills or abilities that generally are or have been learned on the job, though not those that could be acquired in a “brief orientation.” Construct validity refers to a test designed to measure some innate human trait such as honesty. A user of a construct “should show by empirical evidence that the selection procedure is validly related to the construct and that the construct is validly related to the performance of critical or important work behavior(s).”

As a statistical predictive measure, a data mining model could be validated by either criterion-related or construct validity, depending on the trait being sought. Either way, there must be statistical significance showing that the result of the model correlates to the trait (which was already determined to be an important element of job performance). This is an exceedingly low bar for data mining because data mining’s predictions necessarily rest on demonstrated

---

183. 29 C.F.R. §§ 1607.3, 1607.5 (2015). The Guidelines also cite two categories of practices that are unsuitable for validation: informal, unscored practices and technical infeasibility. Id. § 1607.6(B). For the latter case, the Guidelines state that the selection procedure still should be justified somehow or another option should be chosen.

184. Id. § 1607.5(B).

185. Id.

186. Id. § 1607.14(B)(5).

187. Id. §§ 1607.5(F), 1607.14(C).

188. Id. § 1607.14(D)(3).
statistical relationships. Data mining will likely only be used if it is actually predictive of something, so the business necessity defense solely comes down to whether the trait sought is important enough to job performance to justify its use in any context.

Even assuming the Guidelines’ validation requirement is a hurdle for data mining, some courts ignore the Guidelines’ recommendation that an unvalidated procedure be rejected, preferring to rely on “common sense” or finding a “manifest relationship” between the criteria and successful job performance. Moreover, it is possible that the Supreme Court inadvertently overruled the Guidelines in 2009. In Ricci v.Destefano, a case that will be discussed in greater detail in Part III.B, the Court found no genuine dispute that the tests at issue met the job-related and business necessity standards despite not having been validated under the Guidelines and despite the employer actively denying that they could be validated. While the business necessity defense was not directly at issue in Ricci, “[o]n the spectrum between heavier and lighter burdens of justification, the Court came down decidedly in favor of a lighter burden.”

Thus, there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense. Models trained on biased samples, mislabeled examples, and limited features, however, might trigger liability under the alternative employment practice prong. If a plaintiff can show that an alternative, less discriminatory practice that accomplishes the same goals exists and that the employer “refuses” to use it, the employer can be found liable. In this case, a plaintiff could argue that the obvious alternative employment practice would be to fix the problems with the models.

Fixing the models, however, is not a trivial task. For example, in the LinkedIn hypothetical, where the demonstrated interest in different kinds of employees reflects employers’ prejudice, LinkedIn is the party that determines the algorithm by which the discrimination occurs (in this case, based on reacting to third-party preferences). If an employer were to act on the recommendations suggested by the LinkedIn recommendation engine, there

189. Wax, supra note 146, at 633–34.
190. David A. Drachslcer, Assessing the Practical Repercussions of Ricci, AM. CONST. SOC’Y BLOG (July 27, 2009), http://www.acslaw.org/node/13829 [https://perma.cc/AH9G-B3GN] (observing that the Court in Ricci v. DeStefano found no genuine dispute that the unvalidated tests at issue met the job-related and business necessity standards despite the Guidelines creating a presumption of invalidity for unvalidated tests that are discriminatory).
191. New Haven’s primary argument was that it had to withdraw the tests or it would have faced Title VII liability. See Mark S. Brodin, Ricci v. DeStefano: The New Haven Firefighters Case & the Triumph of White Privilege, 20 S. CAL. REV. L. & SOC. JUST. 161, 178 n.128 (2011) (“New Haven forcefully argued throughout the litigation that the exams were ‘flawed’ and may not have identified the most qualified candidates for the supervisory positions.”).
would not be much he could do to make it less reflective of third-party prejudice, aside from calling LinkedIn and asking nicely. Thus, it could not really be said that the employer “refuses” to use an alternative employment practice. The employer could either use the third-party tool or not. Similarly, it might be possible to fix an app like Street Bump that suffers from reporting bias, but the employer would need access to the raw input data in order to do so. In the case of insufficiently rich or granular features, the employer would need to collect more data in order to make the model more discerning. But collecting more data can be time consuming and costly, if not impossible for legal or technical reasons.

Moreover, the under- and overrepresentation of members of protected classes in data is not always evident, nor is the mechanism by which such under- or overrepresentation occurs. The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions. Thus, there are several hurdles to finding disparate impact liability for models employing data that under- or overrepresents members of protected classes. The plaintiff must prove that the employer created or has access to the model, can discover that there is discriminatory effect, and can discover the particular mechanism by which that effect operates. The same can be said for models with insufficiently rich feature sets. Clearly there are times when more features would improve an otherwise discriminatory outcome. But it is, almost by definition, hard to know which features are going to make the model more or less discriminatory. Indeed, it is often impossible to know which features are missing because data miners do not operate with causal relationships in mind. So while theoretically a less discriminatory alternative would almost always exist, proving it would be difficult.

There is yet another hurdle. Neither Congress nor courts have specified what it means for an employer to “refuse” to adopt the less discriminatory procedure. Scholars have suggested that perhaps the employer cannot be held liable until it has considered the alternative and rejected it. Thus, if the employer has run an expensive data collection and analysis operation without ever being made aware of its any discriminatory tendencies, and the employer cannot afford to re-run the entire operation, is the employer “refusing” to use a less discriminatory alternative, or does one simply not exist? How much would the error correction have to cost an employer before it is not seen as a refusal to use the procedure? Should the statute actually be interpreted to mean that an

193. See infra Part III.B.1.
194. See generally Dalessandro, Perlisch & Raeder, supra note 68.
195. Sullivan, supra note 114, at 964; Zimmer, supra note 172, at 505–06.
employer “unreasonably refuses” to use an alternative employment practice? These are all difficult questions, but suffice it to say, the prospect of winning a data mining discrimination case on alternative employment practice grounds seems slim.

The third and final consideration regarding disparate impact liability for data mining is whether a court or Congress might reinvigorate strict business necessity. In that case, things look a little better for plaintiffs bringing disparate impact claims. Where an employer models job tenure, for example, a court may be inclined to hold that it is job related because the model is a “legitimate, non-discriminatory business objective.” But it is clearly not necessary to the job. The same reasoning applies to mining for any single trait that is job related—the practice of data mining is not focused on discovering make-or-break skills. Unless the employer can show that below the cut score, employees cannot do the work, then the strict business necessity defense will fail. Thus, disparate impact that occurs as an artifact of the problem-specification stage can potentially be addressed by strict business necessity.

This reasoning is undermined, though, where employers do not mine for a single trait, but automate their decision process by modeling job performance on a holistic measure of what makes good employees. If employers determine traits of a good employee by simple ratings, and use data mining to appropriately divine good employees’ characteristics among several different variables, then the argument that the model does not account for certain skills that could compensate for the employee’s failings loses its force. Taken to an extreme, an 8,000-feature holistic determination of a “good employee” would still not be strictly “necessary.” Holding a business to such a standard, however, would simply be forbidding that business from ranking candidates if any disparate impact results. Thus, while the strict business necessity defense could prevent myopic employers from creating disparate impacts by their choice of target variable, it would still not address forms of data mining that model general job performance rather than predict specific traits.

Disparate impact doctrine was created to address unintentional discrimination. But it strikes a delicate balance between allowing businesses the leeway to make legitimate business judgments and preventing “artificial, arbitrary, and unnecessary” discrimination. Successful data mining operations will often both predict future job performance and have some

---

197. This would likely require Congressional action because strict business necessity essentially transfers the burden to prove a lack of an alternative employment practice to the defense. By implication, if a practice is “necessary,” there cannot be alternatives. The statute, as it reads now, clearly states that the plaintiff has the burden for that prong. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012).

198. This is an increasingly common practice in low-wage, high-turnover jobs. See Peck, supra note 181.

199. Equal Emp’t Opportunity Comm’n v. Joe’s Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); see also Gallagher v. Magnier, 619 F.3d 823, 834 (8th Cir. 2010).

disparate impact. Unless the plaintiff can find an alternative employment practice to realistically point to, a tie goes to the employer.

C. Masking and Problems of Proof

Masking poses separate problems for finding Title VII liability. As discussed earlier, there is no theoretical problem with finding liability for masking. It is a disparate treatment violation as clear as any. But like traditional forms of intentional discrimination, it suffers from difficulties of proof. While finding intent from stray remarks or other circumstantial evidence is challenging in any scenario, masking presents additional complications for detection.

Data mining allows employers who wish to discriminate on the basis of a protected class to disclaim any knowledge of the protected class in the first instance while simultaneously inferring such details from the data. An employer may want to discriminate by using proxies for protected classes, such as in the case of redlining. Due to housing segregation, neighborhood is a good proxy for race and can be used to redline candidates without reference to race. This is a relatively unsophisticated example, however. It is possible that some combination of musical tastes, stored “likes” on Facebook, and network of friends will reliably predict membership in protected classes. An employer can use these traits to discriminate by setting up future models to sort by these items and then disclaim any knowledge of such proxy manipulation.

More generally, as discussed in Part I, any of the mechanisms by which unintentional discrimination can occur can also be employed intentionally. The example described above is intentional discrimination by proxy, but it is also possible to intentionally bias the data collection process, purposefully mislabel examples, or deliberately use an insufficiently rich set of features, though some of these would probably require a great deal of sophistication. These methods of intentional discrimination will look, for all intents and purposes, identical to the unintentional discrimination that can result from data mining. Therefore, detecting discrimination in the first instance will require the same techniques as detecting unintentional discrimination, namely a disparate impact analysis. Further, assuming there is no circumstantial evidence like an employer’s stray remarks with which to prove intent, a plaintiff might attempt

201. See supra text accompanying notes 106–07.
202. See supra Part I.E.
203. See MASSEY & DENTON, supra note 73, at 51–52.
204. Croll, supra note 88.
207. See Dwork et al., supra note 81, app. at 226 (“Catalog of Evils”).
to prove intent by demonstrating that the employer is using less representative data, poorer examples, or fewer and less granular features than he might otherwise use were he interested in the best possible candidate. That is, one could show that the neutral employment practice is a pretext by demonstrating that there is a more predictive alternative.

This looks like disparate impact analysis. A plaintiff proving masked intentional discrimination asks the same question as in the “alternative employment practice” prong: whether there were more relevant measures the employer could have used. But the business necessity defense is not available in a disparate treatment case, so alternative employment practice is not the appropriate analysis. Scholars have noted, though, that the line between disparate treatment and disparate impact in traditional Title VII cases is not always clear, and sometimes employer actions can be legitimately categorized as either or both. As Professor George Rutherglen has pointed out, “Concrete issues of proof, more than any abstract theory, reveal the fundamental similarity between claims of intentional discrimination and those of disparate impact. The evidence submitted to prove one kind of claim invariably can be used to support the other.” Rutherglen’s point is exactly what must happen in the data mining context: disparate treatment and disparate impact become essentially the same thing from an evidentiary perspective.

To the extent that disparate impact and treatment are, in reality, different theories, they are often confused for each other. Plaintiffs will raise both types of claims as a catch-all because they cannot be sure on which theory they might win, so both theories will be in play in a given case. As a result, courts often seek evidence of state of mind in disparate impact cases and objective, statistical evidence in disparate treatment cases. Assuming the two theories are not functionally the same, using the same evidence for disparate treatment and disparate impact will only lead to more confusion and, as a result, more uncertainty within the courts. Thus, despite its clear nature as a theoretical violation, it is less clear that a plaintiff will be able to win a masking disparate treatment case.

A final point is that traditionally, employers who do not want to discriminate go to great lengths to avoid raising the prospect that they have

211. Rutherglen, supra note 210, at 2320–21.
212. Id. at 2320.
214. Id. at 1153–63.
violated the law. Thus they tend to avoid collecting information about attributes that reveal an individual’s membership in a protected class. Employers even pay third parties to collect relatively easy-to-find information on job applicants, such as professional honors and awards, as well as compromising photos, videos, or membership in online groups, so that the third party can send back a version of the report that “remove[s] references to a person’s religion, race, marital status, disability and other information protected under federal employment laws.” This allows employers to honestly disclaim any knowledge of the protected information. Nonetheless, if an employer seeks to discriminate according to protected classes, she would be able to infer class membership from the data. Thus, employers’ old defense to suspicion of discrimination—that they did not even see the information—is no longer adequate to separate would-be intentional discriminators from employers that do not intend to discriminate.

III.
THE DIFFICULTY FOR REFORMS

While each of the mechanisms for discrimination in data mining presents difficulties for Title VII as currently written, there are also certain obstacles to reforming Title VII to address the resulting problems. Computer scientists and others are working on technical remedies, so to say that there are problems with legal remedies does not suggest that the problems with discrimination in data mining cannot be solved at all. Nonetheless, this Part focuses on the legal aspects. As it illustrates, even assuming that the political will to reform Title VII exists, potential legal solutions are not straightforward.

This Part discusses two types of difficulties with reforming Title VII. First, there are issues internal to the data mining process that make legal reform difficult. For example, the subjectivity in defining a “good employee” is unavoidable, but, at the same time, some answers are clearly less discriminatory than others. How does one draw that line? Can employers gain access to the additional data necessary to correct for collection bias? How much will it cost them to find it? How do we identify the “correct” baseline historical data to avoid reproducing past prejudice or the “correct” level of detail and granularity in a dataset? Before laws can be reformed, policy-level answers to these basic technical, philosophical, and economic questions need to be addressed at least to some degree.

217. For a list of the wide-ranging research underway in computer science, see generally Resources, FATML, http://www.fatml.org/resources.html [https://perma.cc/T2QW-ARHX].
218. See supra Part I.A.
Second, reform will face political and constitutional constraints external to the logic of data mining that will affect how Title VII can be permissibly reformed to address it. Not all of the mechanisms for discrimination seem to be amenable to procedural remedies. If that holds true, only after-the-fact reweighting of results may be able to compensate for the discriminatory outcomes. This is not a matter of missing legislation; it is a matter of practical reality. Unfortunately, while in many cases no procedural remedy will be sufficient, any attempt to design a legislative or judicial remedy premised on reallocation of employment outcomes will not survive long in the current political or constitutional climate, as it raises the specter of affirmative action. Politically, anything that even hints at affirmative action is a nonstarter today, and to the extent that it is permissible to enact such policies, their future constitutionality is in doubt.219

A. Internal Difficulties

I. Defining the Target Variable

Settling on a target variable is a necessarily subjective exercise.220 Disputes over the superiority of competing definitions are often insoluble because the target variables are themselves incommensurable. There are, of course, easier cases, where prejudice or carelessness leads to definitions that subject members of protected classes to avoidably high rates of adverse determinations. But most cases are likely to involve genuine business disagreements over ideal definitions, with each having a potentially greater or lesser impact on protected classes. There is no stable ground upon which to judge the relative merits of definitions because they often reflect competing ideas about the very nature of the problem at issue.221 As Professor Oscar Gandy has argued, “[C]ertain kind[s] of biases are inherent in the selection of the goals or objective functions that automated systems will [be] designed to support.”222 There is no escape from this situation; a target variable must reflect judgments about what really is the problem at issue in making hiring decisions. For certain employers, it might be rather obvious that the problem is one of reducing the administrative costs associated with turnover and training; for others, it might be improving sales; for still others, it might be increasing

219. See Lyle Denniston, Argument Analysis: Now, Three Options on College Affirmative Action, SCOTUSBLOG (Dec. 9, 2015, 2:47 PM), http://www.scotusblog.com/2015/12/argument-analysis-now-three-options-on-college-affirmative-action [https://perma.cc/XF75-N82F] (analysis of oral argument in Fisher v. Univ. of Tex., 758 F.3d 633 (5th Cir. 2014), cert. granted, 135 S. Ct. 2888, (June 29, 2015)); see also Fisher v. Univ. of Tex., 133 S. Ct. 2411, 2419 (2013) (“[A]ny official action that treats a person differently on account of his race or ethnic origin is inherently suspect.” (internal citation omitted)).

220. See supra Part I.A.


innovation. Any argument for the superiority of one target variable over the other will simply make appeals to competing and incommensurate values.

For these same reasons, however, defining the target variable also offers an opportunity for creative thinking about the potentially infinite number of ways of making sound hiring decisions. Data miners can experiment with multiple definitions that each seem to serve the same goal, even if these fall short of what they themselves consider ideal. In principle, employers should rely on proxies that are maximally proximate to the actual skills demanded of the job. While there should be a tight nexus between the sought-after features and these skills, this may not be possible for practical and economic reasons. This leaves data miners in a position to dream up many different nonideal ways to make hiring decisions that may have a greater or less adverse impact on protected classes.

The Second Circuit considered such an approach in Hayden v. County of Nassau. In Hayden, the county’s goal was to find a police entrance exam that was “valid, yet minimized the adverse impact on minority applicants.” The county thus administered an exam with twenty-five parts that could be scored independently. By design, a statistically valid result could be achieved by one of several configurations that counted only a portion of the test sections, without requiring all of them. The county ended up using nine of the sections as a compromise, after rejecting one configuration that was more advantageous to minority applicants but less statistically sound. This is a clear example of defining a problem in such a way that it becomes possible to reduce the disparate impact without compromising the accuracy of the assessment mechanism.

2. Training Data

a. Labeling Examples

Any solution to the problems presented by labeling must be a compromise between a rule that forbids employers from relying on past discrimination and one that allows them to base hiring decisions on historical examples of good employees. In theory, a rule that forbids employers from modeling decisions based on historical examples tainted by prejudice would address the problem of improper labeling. But if the only examples an employer has to draw on are those of past employees who had been subject to discrimination, all learned rules will recapitulate this discrimination.

Title VII has always had to balance its mandate to eliminate discrimination in the workplace with employers’ legitimate discretion. For

223. 180 F.3d 42, 47 (2d Cir. 1999).
224. Id.
225. Id.
226. Id.
example, one of the most common selection procedures that explicitly reproduced past discrimination was seniority. Seniority was, and is still often, a legitimate metric for promotion and is especially important in collective bargaining. After the passage of Title VII, however, seniority was also often used to keep black people from advancing to better jobs because they had not been hired until Title VII forced employers to hire them. Despite this obvious problem with seniority, Title VII contains an explicit carve-out for “bona fide seniority or merit system[s].” As a result, the Supreme Court has held that “absent a discriminatory purpose, the operation of a seniority system cannot be an unlawful employment practice even if the system has some discriminatory consequences.” Given the inherent tension between ensuring that past discrimination is not reproduced in future decisions and permitting employers legitimate discretion, it should be unsurprising that, when translated to data mining, the problem is not amenable to a clear solution.

In fact, this difficulty is even more central to data mining. Data miners who attempt to remove the influence of prejudice on prior decisions by recoding or relabeling examples may find that they cannot easily resolve what the nonprejudicial determination would have been. As Calders and Žliobaitė point out, “[T]he notion of what is the correct label is fuzzy.” Employers are unlikely to have perfectly objective and exhaustive standards for hiring; indeed, part of the hiring process is purposefully subjective. At the same time, employers are unlikely to have discriminated so completely in the past that the only explanation for rejecting an applicant was membership in protected classes. This leaves data miners tasked with correcting for prior prejudice with the impossible challenge of determining what the correct subjective employment decision would have been absent prejudice. Undoing the imprint of prejudice on the data may demand a complete rerendering of the biased decisions rather than simply adjusting those decisions according to some fixed statistical measure.

b. Data Collection

Although there are some cases with obviously skewed datasets that are relatively easy to identify and correct, often the source and degree of the bias will not be immediately apparent. Street Bump suffered from a visually

---

227. Selmi, supra note 178, at 715.
228. See Albemarle Paper Co. v. Moody, 422 U.S. 405, 450 (1975) (Burger, J., concurring) (“The basis of Albemarle’s liability was that its seniority system perpetuated the effects of past discrimination . . . .”).
231. Calders & Žliobaitė, supra note 64, at 48.
232. For example, establishing whether and to what extent crime statistics misrepresent the relative proportion of offenses committed by different social groups is not an easy task. Especially challenging are those crimes that are more likely to go under- or unreported if not directly observed by
evident bias when the data was plotted on a map. Boston’s Office of New Urban Mechanics was therefore able to partner with “a range of academics to take into account issues of equitable access and digital divides.” In many cases, however, an analyst can only determine the extent of—and correct for—unintentional discrimination that results from reporting, sampling, and selection biases if the analyst has access to information that somehow reveals misrepresentations of protected classes in the dataset. Often, there may be no practical alternative method for collecting information that would even reveal the existence of a bias.

Any attempt to correct for collection bias immediately confronts the problem of whether or not the employer recognizes the specific type of bias that is producing disparate results. Then, in order to correct for it, an employer must have access to the underlying data and often an ability to collect more. Where more data is clearly not accessible, data miners can proactively compensate for some of the bias by oversampling underrepresented communities.

If the employer fails to be proactive or tries and fails to detect the bias that causes the disparate impact, liability is an open question. As discussed in Part II.B, liability partly depends on how liberally a court interprets the requirement that an employer “refuses” to use an alternative scheme. Even a liberal interpretation, though, would require evidence of the particular type of discrimination at issue, coupled with evidence that such an alternative scheme exists. Thus, finding liability seems unlikely. Worse, where such showing is possible, there may be no easy or obvious way to remedy the situation.

To address collection bias directly, an employer or an auditor must have access to the underlying data and the ability to adjust the model. Congress could require this directly of any employer using data mining techniques. Some employers are investing in their own data now and could potentially meet such requirements. But employers also seem happy to rely on models developed and administered by third parties, who may have a far greater set of examples and far richer data than any individual company. Furthermore, due to economies of scale that are especially important in data analysis, one can imagine that third parties specializing in work-force science will be able to offer employers this service much less expensively than they could manage it.

---


235. See supra Part II.B.

236. Peck, supra note 181.

237. See Richtel, supra note 69.
themselves. If Congress attempted to demand that employers have access to the data, it would face strong resistance from the ever-growing data analysis industry, whose business depends on the proprietary nature of the amassed information. More likely, Congress could require audits by a third party like the EEOC or a private auditor, in order to protect trade secrets, but this still seems a tall task. Ultimately, because proactive oversampling and retroactive data correction are at least possible, collection bias has the most promising prospects for a workable remedy of any of the identified data mining mechanisms.

3. Feature Selection

Even in the absence of prejudice or bias, determining the proper degree of precision in the distinctions drawn through data mining can be extremely difficult. Under formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds. It is far less clear, however, what constitutes legitimate statistical discrimination when individuation does not rely on proscribed criteria. In these cases, the perceived legitimacy seems to depend on a number of factors: (1) whether the errors seem avoidable because (2) gaining access to additional or more granular data would be trivial or (3) would not involve costs that (4) outweigh the benefits. This seems to suggest that the task of evaluating the legitimacy of feature selection can be reduced to a rather straightforward cost-benefit analysis. Companies would have an obligation to pursue ever more—and more granular—data until the costs of gathering that data exceed the benefits conferred by the marginal improvements in accuracy.

Unfortunately, as is often the case with cost-benefit analyses, this approach fails to consider how different actors will perceive the value of the supposed benefits as well as the costs associated with errors. The obvious version of this criticism is that “actuarially saddled” victims of inaccurate determinations may find cold comfort in the fact that certain decisions are rendered more reliably overall when decision makers employ data mining. A more sophisticated version of this criticism focuses on the way such errors assign costs and benefits to different actors at systematically different rates. A model with any error rate that continues to turn a profit may be acceptable to decision makers at a company, no matter the costs or inconvenience to specific customers. Even when companies are subject to market pressures that would

---

238. Schauer, supra note 67, at 5. As Schauer explains, perfectly particularized decisions are, of course, a logical impossibility. Accepting this inherent limitation introduces a different sort of procedural concern: occasional errors might be tolerable if they are easy to detect and rectify, which is why, among other things, the perceived legitimacy of decisions often also depends on due process. See id. at 172; see also Citron, supra note 11.

239. Gandy, supra note 31, at 36.
force them to compete by lowering these error rates, the companies may find that there is simply no reason to invest in efforts that do so if the errors happen to fall disproportionately on especially unprofitable groups of consumers. Furthermore, assessing data mining as a matter of balancing costs and benefits leaves no room to consider morally salient disparities in the degree to which the costs are borne by different social groups. This raises the prospect that there might be systematic differences in the rates at which members of protected classes are subject to erroneous determinations. Condemning these groups to bear the disproportionate burden of erroneous determinations would strike many as highly objectionable, despite greater accuracy in decision making for the majority group. Indeed, simply accepting these cost differences as a given would subject those already in less favorable circumstances to less accurate determinations.

Even if companies assume the responsibility for ensuring that members of protected classes do not fall victim to erroneous determinations at systematically higher rates, they could find that increasing the resolution and range of their analyses still fails to capture the causal relationships that account for different outcomes because those relationships are not easily represented in data. In such cases, rather than reducing the error rate for those in protected classes, data miners could structure their analyses to minimize the difference in error rates between groups. This solution may involve some unattractive tradeoffs, however. In reducing the disparate impact of errors, it may increase the overall amount of errors. In other words, generating a model that is equally unfair to protected and unprotected classes might increase the overall amount of unfairness.

4. Proxies

Computer scientists have been unsure how to deal with redundant encodings in datasets. Simply withholding these variables from the data mining exercise often removes criteria that hold demonstrable and justifiable relevance to the decision at hand. As Calders and Žliobaitė note, “[I]t is problematic [to remove a correlated attribute] if the attribute to be removed also carries some objective information about the label [quality of interest].” Part of the problem seems to be that there is no obvious way to determine how correlated a relevant attribute must be with class membership to be worrisome. Nor is there a self-evident way to determine when an attribute is sufficiently relevant to justify its consideration, despite its high correlation with class membership. As

241. See, e.g., Gandy, supra note 31, at 39.
242. See supra note 64 and accompanying text.
243. Calders & Žliobaitė, supra note 64, at 54.
Professors Devin Pope and Justin Sydnor explain, “[V]ariables are likely neither solely predictive nor purely proxies for omitted characteristics.”

But there is a bigger problem here: attempting to ensure fairly rendered decisions by excising highly correlated criteria only makes sense if the disparate impact happens to be an avoidable artifact of a particular way of rendering decisions. And yet, even when denied access to these highly correlated criteria, data mining may suggest alternative methods for rendering decisions that still result in the same disparate impact. Focusing on isolated data points may be a mistake because class membership can be encoded in more than one specific and highly correlated criterion. Indeed, it is very likely that class membership is reflected across a number of interrelated data points. But such outcomes might instead demonstrate something more unsettling: that other relevant criteria, whatever they are, happen to be possessed at different rates by members of protected classes. This explains why, for instance, champions of predictive policing have responded to critics by arguing that “[i]f you wanted to remove everything correlated with race, you couldn’t use anything. That’s the reality of life in America.”

Making accurate determinations means considering factors that are somehow correlated with proscribed features.

Computer scientists have even shown that “[r]emoving all such correlated attributes before training does remove discrimination, but with a high cost in classifier accuracy.” This reveals a rather uncomfortable truth: the current distribution of relevant attributes—attributes that can and should be taken into consideration in apportioning opportunities fairly—is demonstrably correlated with sensitive attributes because the sensitive attributes have meaningfully conditioned what relevant attributes individuals happen to possess. As such, attempts to ensure procedural fairness by excluding certain criteria from consideration may conflict with the imperative to ensure accurate determinations. The only way to ensure that decisions do not systematically disadvantage members of protected classes is to reduce the overall accuracy of

---

245. Supra discussion accompanying note 101.
246. Labi, supra note 5 (quoting Ellen Kurtz, Director of Research for Philadelphia’s Adult Probation and Parole Department).
248. In a sense, computer scientists have unwittingly furnished the kind of evidence that social scientists routinely seek: the particular contours of inequality. See, e.g., SOCIAL INEQUALITY (Kathryn M. Neckerman ed., 2004).
all determinations. As Dwork et al. remark, these results “demonstrate a quanti[t]ative trade-off between fairness and utility.”249

In certain contexts, data miners will never be able to fully disentangle legitimate and proscribed criteria. For example, the workforce optimization consultancy, Evolv, discovered that “[d]istance between home and work . . . is strongly associated with employee engagement and retention.”250 Despite the strength of this finding, Evolv “never factor[s] [it] into the score given each applicant . . . because different neighborhoods and towns can have different racial profiles, which means that scoring distance from work could violate equal-employment-opportunity standards.”251

Scholars have taken these cases as a sign that the “major challenge is how to find out which part of information carried by a sensitive (or correlated) attribute is sensitive and which is objective.”252 While researchers are well aware that this may not be easy to resolve, let alone formalize into a computable problem, there is a bigger challenge from a legal perspective: any such undertaking would necessarily wade into the highly charged debate over the degree to which the relatively less favorable position of protected classes warrants the protection of antidiscrimination law in the first instance.

The problems that render data mining discriminatory are very rarely amenable to obvious, complete, or welcome resolution. When it comes to setting a target variable and feature selection, policy cannot lay out a clear path to improvement; reducing the disparate impact will necessitate open-ended exploration without any way of knowing when analysts have exhausted the possibility for improvement. Likewise, policies that compel institutions to correct tainted datasets or biased samples will make impossible demands of analysts. In most cases, they will not be able to determine what the objective determination should have been or independently observe the makeup of the entire population. Dealing with both of these problems will ultimately fall to analysts’ considered judgment. Solutions that reduce the accuracy of decisions to minimize the disparate impact caused by coarse features and unintentional proxies will force analysts to make difficult and legally contestable trade-offs. General policies will struggle to offer the specific guidance necessary to determine the appropriate application of these imperfect solutions. And even when companies voluntarily adopt such strategies, these internal difficulties will likely allow a disparate impact to persist.

249. Dwork et al., supra note 81, at 215; cf. Wax, supra note 146, at 711 (noting intractable problems due to a “validity-diversity tradeoff” in employment metrics).
250. Peck, supra note 181.
251. Id. Other companies have not held back from considering this information for the very same purposes. See Joseph Walker, Meet the New Boss: Big Data, WALL ST. J. (Sept. 20, 2012), http://www.wsj.com/news/articles/SB1000087239639044389030457800625019616768 [https://perma.cc/6DHY-M429].
252. Calders & Žliobaité, supra note 64, at 56.
B. External Difficulties

Assuming the internal difficulties can be resolved, there are further political and constitutional restraints on addressing Title VII’s inadequacies with respect to data mining. Data mining discrimination will force a confrontation between the two divergent principles underlying antidiscrimination law: anticlassification and antisubordination. Which of these two principles motivates discrimination law is a contentious debate, and making remedies available under antidiscrimination law will require a commitment to antisubordination principles that have thus far not been forthcoming from legislatures. This is not merely a political concern, as substantive remediation is becoming ever more suspect constitutionally as well. While such remedies may be politically and legally impossible, the nature of data mining itself makes them practically necessary. Accordingly, these external difficulties may prevent antidiscrimination law from fully addressing data mining discrimination.

Two competing principles have always undergirded antidiscrimination law: anticlassification and antisubordination. Anticlassification is the narrower of the two, holding that the responsibility of the law is to eliminate the unfairness individuals in certain protected classes experience due to decision makers’ choices. Antisubordination theory, in contrast, holds that the goal of antidiscrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance.

Different mitigation policies effectuate different rationales. Disparate treatment doctrine arose first, clearly aligning with the anticlassification principle by proscribing intentional discrimination, in the form of either explicit singling out of protected classes for harm or masked intentional discrimination. Since disparate impact developed, however, there has never been clarity as to which of the principles it is designed to effectuate. On the one hand, disparate impact doctrine serves anticlassification by being an “evidentiary dragnet” used to “smoke out” well-hidden disparate treatment. On the other hand, as an effects-based doctrine, there is good reason to believe it was intended to address substantive inequality. In this sense, the “business


254. See Norton, supra note 253.

255. Id. at 209.

256. Id. at 206.

257. Primus, supra note 98, at 520–23.

258. Id.; Perry, supra note 138, at 526.

259. See Griggs v. Duke Power Co., 401 U.S. 424, 429–30 (1971) (“The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of
necessity” defense is a necessary backstop that prevents members of traditionally disadvantaged groups from simply forcing their way in without the necessary skills or abilities.  

Thus, the mapping from anticlassification and antisubordination to disparate treatment and disparate impact was never clean. Early critics of civil rights laws actually complained that proscribing consideration of protected class was a subsidy to black people. This argument quickly gave way in the face of the rising importance of the anticlassification norm. Over the years, the anticlassification principle has come to dominate the landscape so thoroughly that a portion of the populace thinks (as do a few Justices on the Supreme Court) that it is the only valid rationale for antidiscrimination law.  

The move away from antisubordination began only five years after disparate impact was established in Griggs. In Washington v. Davis, the Court held that disparate impact could not apply to constitutional claims because equal protection only prohibited intentional discrimination. Since then, the various affirmative action cases have overwritten the distinction between benign and harmful categorizations of race in favor of a formalistic anticlassification principle, removed from its origins as a tool to help members of historically disadvantaged groups. White men can now bring disparate treatment claims. If antidiscrimination law is no longer thought to serve the purpose of improving the relative conditions of traditionally disadvantaged groups, antisubordination is not part of the equation. 

While the Court has clearly established that antisubordination is not part of constitutional equal protection doctrine, that it does not mean that antisubordination cannot animate statutory antidiscrimination law. 

Antisubordination and anticlassification came into sharp conflict in Ricci v. DeStefano, a 2009 case in which the City of New Haven refused to certify a promotion exam given to its firefighters on the grounds that it would have produced a disparate impact based on its results. The Supreme Court held that the refusal to certify the test, a facially race-neutral attempt to correct for perceived disparate impact, was in fact a race-conscious remedy that constituted disparate treatment of the majority-white firefighters who would

employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to ‘freeze’ the status quo of prior discriminatory employment practices.

262. Id.
263. See Bagenstos, supra note 122, at 41.
267. Id.
have been promoted based on the exam’s results.\textsuperscript{268} The Court held that disparate treatment cannot be a remedy for disparate impact without a “strong basis in evidence” that the results would lead to actual disparate treatment liability.\textsuperscript{269}

\textit{Ricci} was the first indication at the Supreme Court that disparate impact doctrine could be in conflict with disparate treatment.\textsuperscript{270} The Court had previously ruled in essence that the antisubordination principle could not motivate a constitutional decision,\textsuperscript{271} but it had not suggested that law effectuating that principle could itself be discriminatory against the dominant groups. That has now changed.\textsuperscript{272}

The decision has two main consequences for data mining. First, where the internal difficulties in resolving discrimination in data mining described above can be overcome, legislation that requires or enables such resolution may run afoul of \textit{Ricci}. Suppose, for example, Congress amended Title VII to require that employers make their training data and models auditable. In order to correct for detected biases in the training data that result in a model with a disparate impact, the employer would first have to consider membership in the protected class. The remedy is inherently race-conscious. The \textit{Ricci} Court did hold that an employer may tweak a test during the “test-design stage,” however.\textsuperscript{273} So, as a matter of timing, data mining might not formally run into

\begin{footnotes}
\footnote{268. \textit{Id.}}
\footnote{269. \textit{Id.} at 563.}
\footnote{270. Primus, supra note 92, at 1344; Lawrence Rosenthal, Saving Disparate Impact, 34 \textit{Cardozo L. Rev.} 2157, 2162–63 (2013); Norton, supra note 253, at 229.}
\footnote{271. See Washington v. Davis, 426 U.S. 229, 239 (holding that discriminatory purpose is necessary to finding a violation of equal protection).}
\footnote{272. Primus, supra note 92, at 1343. While the decision was formally about Title VII only, and thus amenable to statutory resolution, the reasoning applied equally well to a future equal protection claim, endangering the future of disparate impact. \textit{Id.} at 1385–87; Bradley A. Areheart, The Anticlassification Turn in Employment Discrimination Law, 63 \textit{A.L.A. L. Rev.} 955, 994 (2012); Norton, supra note 253, at 229–30. Justice Scalia stated as much in his concurrence. \textit{Ricci}, 557 U.S. at 594 (Scalia, J., concurring) ("[The Court’s] resolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution’s guarantee of equal protection?"). But the Supreme Court seemed to pull back from the brink last term, approving of the use of disparate impact in a new setting—the Fair Housing Act—and engaging deeply with the constitutional issues that \textit{Ricci} raised, settling them for now. Samuel R. Bagenstos, Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities, 101 \textit{Cornell L. Rev.}, at *11–12 (forthcoming 2016), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642631 [https://perma.cc/WD43-XW2G]; Richard Primus, Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After \textit{Ricci} and Inclusive Communities, in \textit{Title VII of the Civil Rights Act After 50 Years: Proceedings of the New York University 67th Annual Conference on Labor} 295 (2015).}
\footnote{273. \textit{Ricci}, 557 U.S. at 585 (majority opinion) ("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end.").}
Ricci if the bias resulting in a disparate impact is corrected before applied to individual candidates. After an employer begins to use the model to make hiring decisions, only a “strong basis in evidence” that the employer will be successfully sued for disparate impact will permit corrective action.274 Of course, unless every single model used by an employer is subject to a prescreening audit (an idea that seems so resource intensive that it is effectively impossible), the disparate impact will be discovered only when the employer faces complaints. Additionally, while Ricci’s holding was limited in scope, the “strong basis in evidence” standard did not seem to be dictated by the logic of the opinion, which illustrated a more general conflict between disparate treatment and disparate impact.275

Second, where the internal difficulties cannot be overcome, there is likely no way to correct for the discriminatory outcomes aside from results-focused balancing, and requiring this will pose constitutional problems. For those who adhere to the anticlassification principle alone, such an impasse may be perfectly acceptable. They might say that as long as employers are not intentionally discriminating based on explicitly proscribed criteria, the chips should fall where they may. To those who believe some measure of substantive equality is important over and above procedural equality, this result will be deeply unsatisfying.

An answer to the impasse created by situations that would require results-focused rebalancing is to reexamine the purpose of antidiscrimination law. The major justification for reliance on formal disparate treatment is that prejudice is simply irrational and thus unfair. But if an employer knows that his model has a disparate impact, but it is also his most predictive, the argument that the discrimination is irrational loses any force. Thus, data mining may require us to reevaluate why and whether we care about not discriminating.

Consider another example involving tenure predictions, one in which an employer ranks potential employees with the goal of hiring only those applicants that the company expects to retain for longer periods of time. In optimizing its selection of applicants in this manner, the employer may unknowingly discriminate against women if the historical data demonstrates that they leave their positions after fewer years than their male counterparts. If gender accounts for a sufficiently significant difference in employee tenure, data mining will generate a model that simply discriminates on the basis of gender or those criteria that happen to be proxies for gender. Although selecting applicants with an eye to retention might seem both rational and reasonable, granting significance to predicted tenure would subject women to systematic disadvantage if gender accounts for a good deal of the difference in tenure. If that is the case, any data mining exercise that attempts to predict

274. Id. at 585.
275. See generally id.
tenure will invariably rediscover this relationship. One solution could be for Congress to amend Title VII to reinvigorate strict business necessity. This would allow a court to accept that relying on tenure is rational but not strictly “necessary” and that perhaps other factors could make up for the lack of predicted tenure.

But this solution and all others must rely on the antisubordination principle. Consider this question: should the law permit a company to hire no women at all—or none that it correctly predicts will depart following the birth of a child—because it is the most rational choice according to their model? The answer seems obviously to be no. But why not? What forms the basis for law’s objection to rational decisions, based on seemingly legitimate criteria, that place members of protected classes at systematic disadvantage? The Supreme Court has observed that, “Title VII requires employers to treat their employees as individuals, not ‘as simply components of a racial, religious, sexual, or national class.’” On the strength of that statement, the Court held that employers could not force women to pay more into an annuity because they, as women, were likely to live longer. But it is not clear that this reasoning translates directly to data mining. Here, the model takes a great deal of data about an individual, and while it does make a determination based on statistics, it will make a different one if analyzing two different women. So if the model said to hire no women, it would be illegal, but, according to the doctrine, perhaps only because every woman ends up with the same result.

The only escape from this situation may be one in which the relevance of gender in the model is purposefully ignored and all factors correlated with gender are suppressed. The outcome would be a necessarily less accurate model. The justification for placing restrictions on employers, and limiting the effectiveness of their data mining, would have to depend on an entirely different set of arguments than those advanced to explain the wrongfulness of biased data collection, poorly labeled examples, or an impoverished set of features. Here, shielding members of protected classes from less favorable treatment is not justified by combatting prejudice or stereotyping. In other words, any prohibition in this case could not rest on a procedural commitment to ensuring ever more accurate determinations. Instead, the prohibition would have to rest on a substantive commitment to equal representation of women in the workplace. That is, it would have to rest on a principle of antisubordination.

276. Remember that if there is disparate impact, but no liability, it is because the goal was deemed job-related or satisfied business necessity.
277. As a matter of case law, this question has essentially been answered. The Supreme Court has ruled that in the case of women being required to pay more into an annuity because they would likely live longer, pure market rationality is not a good enough answer. Ariz. Governing Comm. v. Norris, 463 U.S. 1073, 1083 (1983) (quoting City of Los Angeles Dep’t of Water & Power v. Manhart, 435 U.S. 702, 708 (1978)).
278. Id.
279. Id.
The dilemma is clear: the farther the doctrine gets from substantive remediation, the less utility it has in remedying these kinds of discriminatory effects. But the more disparate impact is thought to embody the antisubordination principle—as opposed to the “evidentiary dragnet” in service of the anticlassification norm—the more it will invite future constitutional challenges.

This also raises a point about disparate treatment and data mining. Within data mining, the effectiveness of prohibiting the use of certain information exists on a spectrum. On one end, the prohibition has little to no effect because either the information is redundantly encoded or the results do not vary along lines of protected class. On the other end, the prohibition reduces the accuracy of the models. That is, if protected class data were not prohibited, that information would alter the results, presumably by making members of protected classes worse (or, in some cases, better) off. Thus, as a natural consequence of data mining, a command to ignore certain data has either no effect or the effect of altering the fortunes of those protected classes in substantive ways. Therefore, with respect to data mining, due to the zero-sum nature of a ranking system, even disparate treatment doctrine is a reallocative remedy similar to affirmative action. Once again, this erodes the legitimate rationale for on the one hand supporting an anticlassification principle but on the other, holding fast against antisubordination in this context. The two principles tend to accomplish the same thing, but one is less effective at achieving substantive equality.

This reveals that the pressing challenge does not lie with ensuring procedural fairness through a more thorough stamping out of prejudice and bias but rather with developing ways of reasoning to adjudicate when and what amount of disparate impact is tolerable. Abandoning a belief in the efficacy of procedural solutions leaves policy makers in an awkward position because there is no definite or consensus answer to questions about the fairness of specific outcomes. These need to be worked out on the basis of different normative principles. At some point, society will be forced to acknowledge that this is really a discussion about what constitutes a tolerable level of disparate impact in employment. Under the current constitutional order and in the political climate, it is tough to even imagine having such a conversation. But, until that happens, data mining will be permitted to exacerbate existing inequalities in difficult-to-counter ways.

280. Id. at 537.
282. See supra text accompanying note 101.
283. For an argument that this is true more generally, see Bagenstos, supra note 90, and Owen M. Fiss, A Theory of Fair Employment Laws, 38 U. CHI. L. REV. 235, 313 (1971) (arguing that a key to understanding antidiscrimination prohibitions in the employment realm is that the prohibitions “confer[] benefits on a racial class—blacks”).
CONCLUSION

This Essay has identified two types of discriminatory outcomes from data mining: a family of outcomes where data mining goes “wrong” and outcomes where it goes too “right.” Data mining can go wrong in any number of ways. It can choose a target variable that correlates to protected class more than others would, reproduce the prejudice exhibited in the training examples, draw adverse lessons about protected classes from an unrepresentative sample, choose too small a feature set, or not dive deep enough into each feature. Each of these potential errors is marked by two facts: the errors may generate a manifest disparate impact, and they may be the result of entirely innocent choices made by data miners.

Where data mining goes “right,” data miners could not have been any more accurate given the starting point of the process. This very accuracy, exposing an uneven distribution of attributes that predict the target variable, gives such a result its disparate impact. If the data accurately models inequality, attempts to devise an alternative way of making the same prediction will only narrow the disparate impact if these efforts reduce the accuracy of the decision procedure. By now, it should be clear that Title VII, and very likely other similarly process-oriented civil rights laws, cannot effectively address this situation.

This means something different for the two families, and it should be slightly more surprising for the former. At a high level of abstraction, where a decision process goes “wrong” and this wrongness creates a disparate impact, Title VII and similar civil rights laws should be up to the task of solving the problem; that is ostensibly their entire purpose. But aside from a few more obvious cases involving manifest biases in the dataset, it is quite difficult to determine ahead of time what “correct” data mining looks like. A decision maker can rarely discover that the choice of a particular target variable is more discriminatory than other choices until after the fact, at which point it may be difficult and costly to change course. While data miners might have some intuitions about the influence that prejudice or bias played in the prior decisions that will serve as training data, data miners may not have any systematic way of measuring and correcting for that influence. And even though ensuring reliable samples before training a model is a possibility, the data may never be perfect. It may be impossible to determine, ex ante, how much the bias contributes to the disparate impact, it may not be obvious how to collect additional data that makes the sample more representative, and it may be prohibitively expensive to do so. Companies will rarely be able to resolve these problems completely; their models will almost always suffer from some deficiency that results in a disparate impact. A standard that holds companies liable for any amount of theoretically avoidable disparate impact is likely to ensnare all companies. Thus, even at this level of abstraction, it becomes clear that holding the decision makers responsible for these disparate impacts is at
least partly troubling from a due process perspective. Such concerns may counsel against using data mining altogether. This would be a perverse outcome, given how much even imperfect data mining can do to help reduce the very high rates of discrimination in employment decisions.

If liability for getting things “wrong” is difficult to imagine, how does liability for getting things “right” make any more sense? That proxy discrimination largely rediscovers preexisting inequalities suggests that perhaps Title VII is not the appropriate remedial vehicle. If what is at stake are the results of decades of historical discrimination and wealth concentration that have created profound inequality in society, is that not too big a problem to remedy through individual lawsuits, assuming affirmative action and similar policies are off the table? Thus, perfect data mining forces the question: if employers can say with certainty that, given the status quo, candidates from protected classes are on average less ready for certain jobs than more privileged candidates, should employers specifically be penalized for hiring fewer candidates from protected classes?

Doctrinally, the answer is yes, to some extent. Professor Christine Jolls has written that disparate impact doctrine is akin to accommodation in disability law—that is, both accommodations and disparate impact specifically require employers to depart from pure market rationality and incur costs associated with employing members of protected classes. Similarly, the Title VII annuity cases and Title VII’s ban on following racist third-party preferences each require a departure from market rationality. Thus, Title VII makes that decision to a degree. But to what degree? How much cost must an employer bear?

Title VII does not require an employer to use the least discriminatory means of running a business. Likewise, Title VII does not aim to remedy historical discrimination and current inequality by imposing all the costs of restitution and redistribution on individual employers. It is more appropriately understood as a standard of defensible disparate impact. One route, then, to addressing the problems is to make the inquiry more searching and put the burden on the employer to avoid at least the easy cases. In a system that is as unpredictable as data mining can be, perhaps the proper way of

---

284. We cannot stress enough the import of these caveats. Certainty is a strong and unlikely precondition, and the status quo should not be taken as a given, as we explain below.
285. See generally Jolls, supra note 90.
287. See 29 C.F.R. § 1604.2(a)(1)(iii) (2015) (stating the EEOC’s position that “the preferences of coworkers, the employer, clients or customers” cannot be used to justify disparate treatment).
thinking about the solution is a duty of care, a theory of negligent discrimination.290

But if Title VII alone cannot solve these problems, where should society look for answers? Well, the first answer is to question the status quo. Data mining takes the existing state of the world as a given and ranks candidates according to their predicted attributes in that world. Data mining, by its very nature, treats the target variable as the only item that employers are in a position to alter; everything else that happens to correlate with different values for the target variable is assumed stable. But there are many reasons to question these background conditions. Sorting and selecting individuals according to their apparent qualities hides the fact that the predicted effect of possessing these qualities with respect to a specific outcome is also a function of the conditions under which these decisions are made. Recall the tenure example from Part III.B. In approaching appropriate hiring practices as a matter of selecting the “right” candidates at the outset, an employer will fail to recognize potential changes that he could make to workplace conditions. A more family-friendly workplace, greater on-the-job training, or a workplace culture more welcoming to historically underrepresented groups could affect the course of employees’ tenure and their long-term success in ways that undermine the seemingly prophetic nature of data mining’s predictions.

These are all traditional goals for reducing discrimination within the workplace, and they continue to matter even in the face of the eventual widespread adoption of data mining. But data can play a role here, too. For example, comparing the performance of equally qualified candidates across different workplaces can help isolate the formal policies and institutional dynamics that are more or less likely to help workers flourish. Research of this sort could also reveal areas for potential reform.291

Education is also important. Employers may take some steps to rectify the problem on their own if they better understand the cause of the disparity. Right now, many of the problems described in Part I are relatively unknown. But the more employers and data miners understand these pitfalls, the more they can strive to create better models on their own. Many employers switch to data-driven practices for the express purpose of eradicating bias;292 if employers discover that they are introducing new forms of bias, they can correct course.

Even employers seeking only to increase efficiency or profit may find that their incentives align with the goals of nondiscrimination. Faulty data and data

mining will lead employers to overlook or otherwise discount people who are actually “good” employees. Where the cost of addressing these problems is at least compensated for by a business benefit of equal or greater value, employers may have natural incentives to do so.

Finally, employers could also make more effective use of the tools that computer scientists have begun to develop.\footnote{293} Advances in these areas will depend, crucially, on greater and more effective collaboration between employers, computer scientists, lawyers, advocates, regulators, and policy makers.\footnote{294}

This Essay is a call for caution in the use of data mining, not its abandonment. While far from a panacea, data mining can and should be part of a panoply of strategies for combatting discrimination in the workplace and for promoting fair treatment and equality. Ideally, institutions can find ways to use data mining to generate new knowledge and improve decision making that serves the interests of both decision makers and protected classes. But where data mining is adopted and applied without care, it poses serious risks of reproducing many of the same troubling dynamics that have allowed discrimination to persist in society, even in the absence of conscious prejudice.

\footnote{293}{See list supra note 217.}
Losing Out on Employment Because of Big Data Mining

Solon Barocas is a postdoctoral research associate at the Center for Information Technology Policy at Princeton University. Andrew Selbst is the Alan Morrison Supreme Court Assistance fellow at Public Citizen.

August 6, 2014

Unfortunately, our civil rights statutes may not be up to the task of addressing the discriminatory effects of commercial data mining. Antidiscrimination law is primarily focused on prejudice, but there is an even greater risk that seemingly innocuous decisions by data miners could result in unintentional discrimination that can be just as severe.

When companies use patterns in large datasets to hire employees, they may unknowingly rely on previous poor decisions.

For example, companies may attempt to improve their hiring decisions by looking for patterns in the online behavior of prior applicants who later performed well as employees. While this has the possibility to suggest new and surprising criteria for job candidates, the company can only draw lessons from the limited set of employees that it has hired in the past.

Worse, data mining can pick up the bad examples set by previous hiring decisions: If past prejudice denied certain classes of candidates the opportunity to demonstrate their talents, data mining may fail to find anything to like in many qualified individuals who happen to fall into that category.

It is also not always evident where predictions are skewed by selection bias or past prejudice. Even where these are recognized as potential problems, the appropriate solution is not always clear. How do employers account for the kinds of candidates they have never hired in the past? What corrective measures can they take to purge the often subtle influence of prejudice or implicit bias from its hiring records? And how much should employers be expected to bear the costs that these solutions would entail?

It is also not clear that richer data sets — even those untainted by bias or prejudice — would be more beneficial to disadvantaged groups. Data mining may still have a disproportionately negative effect on protected classes if the criteria that reliably predict some job-related quality also happen to correlate with class membership. This could further entrench inequality for protected classes.

Liability in these cases is unclear. The disparate impact doctrine — which deals with "practices that are fair in form, but discriminatory in operation" — contains an exception that allows employers to make decisions necessary to their businesses. Where data mining ranks candidates outright and involves no errors, but still evidences a disparate impact, it may fall into this exception, because an employer using the system could believe he was merely choosing the best candidate. The exception may apply even though another way of mining the data might rank candidates differently.

How much onus should we place on employers to explore alternatives in these situations? Discrimination in data mining will eventually force us to examine or reexamine several tough, but basic, questions about fairness.
Imagine if, during the Jim Crow era, a newspaper offered advertisers the option of placing ads only in copies that went to white readers.

That’s basically what Facebook is doing nowadays.

The ubiquitous social network not only allows advertisers to target users by their interests or background, it also gives advertisers the ability to exclude specific groups it calls “Ethnic Affinities.” Ads that exclude people based on race, gender and other sensitive factors are prohibited by federal law in housing and employment.

Here is a screenshot of an ad we purchased in Facebook’s housing categories via the company’s advertising portal:
Facebook’s racial exclusion options to a prominent civil rights lawyer John Relman, he gasped and said, “This is horrifying. This is massively illegal. This is about as blatant a violation of the federal Fair Housing Act as one can find.”

The Fair Housing Act of 1968 makes it illegal "to make, print, or publish, or cause to be made, printed, or published any notice, statement, or advertisement, with respect to the sale or rental of a dwelling that indicates any preference, limitation, or discrimination based on race, color, religion, sex, handicap, familial status, or national origin.” Violators can face tens of thousands of dollars in fines.

The Civil Rights Act of 1964 also prohibits the “printing or publication of notices or advertisements indicating prohibited preference, limitation, specification or discrimination” in employment recruitment.

Facebook’s business model is based on allowing advertisers to target specific groups — or, apparently to exclude specific groups — using huge reams of personal data the company has collected about its users. Facebook’s microtargeting is particularly helpful for advertisers looking to reach niche audiences, such as swing-state voters concerned about climate change. ProPublica recently offered a tool allowing users to see how Facebook is
categorizing them. We found nearly 50,000 unique categories in which Facebook places its users.

Facebook says its policies prohibit advertisers from using the targeting options for discrimination, harassment, disparagement or predatory advertising practices.

“We take a strong stand against advertisers misusing our platform: Our policies prohibit using our targeting options to discriminate, and they require compliance with the law,” said Steve Satterfield, privacy and public policy manager at Facebook. “We take prompt enforcement action when we determine that ads violate our policies.”

Satterfield said it’s important for advertisers to have the ability to both include and exclude groups as they test how their marketing performs. For instance, he said, an advertiser “might run one campaign in English that excludes the Hispanic affinity group to see how well the campaign performs against running that ad campaign in Spanish. This is a common practice in the industry.”

He said Facebook began offering the “Ethnic Affinity” categories within the past two years as part of a “multicultural advertising” effort.

Satterfield added that the “Ethnic Affinity” is not the same as race — which Facebook does not ask its members about. Facebook assigns members an “Ethnic Affinity” based on pages and posts they have liked or engaged with on Facebook.

When we asked why “Ethnic Affinity” was included in the “Demographics” category of its ad-targeting tool if it’s not a representation of demographics, Facebook responded that it plans to move “Ethnic Affinity” to another section.

Facebook declined to answer questions about why our housing-categories ad excluding minority groups was approved 15 minutes after we placed the order.

Breaking the Black Box
We live in an era of increasing automation. But as machines make more decisions for us, it is increasingly important to understand the algorithms that produce their judgments. See the series.

Machine Bias
We’re investigating algorithmic injustice and the formulas that increasingly influence our
New York Times has put in place to prevent discriminatory housing ads.

After the newspaper was successfully sued under the Fair Housing Act in 1989, it agreed to review ads for potentially discriminatory content before accepting them for publication.

Steph Jespersen, the Times’ director of advertising acceptability, said that the company’s staff runs automated programs to make sure that ads that contain discriminatory phrases such as “whites only” and “no kids” are rejected.

The Times’ automated program also highlights ads that contain potentially discriminatory code words such as “near churches” or “close to a country club.” Humans then review those ads before they can be approved.

Jespersen said the Times also rejects housing ads that contain photographs of too many white people. The people in the ads must represent the diversity of the population of New York, and if they don’t, he says he will call up the advertiser and ask them to submit an ad with a more diverse lineup of models.

But, Jespersen said, these days most advertisers know not to submit discriminatory ads: “I haven’t seen an ad with ‘whites only’ for a long time.”

**Clarification, Oct. 28, 2016:** We’ve updated the story to explain more clearly that the ad we bought was not for housing itself — it was placed in Facebook’s housing categories.
A new study says that Facebook’s ad delivery algorithm discriminates based on race and gender, even when advertisers are trying to reach a broad audience. The research backs up a similar claim that the US Department of Housing and Urban Development made last week when it sued Facebook for breaking housing discrimination laws. It also expands the scope of an already potentially damning body of research about online advertising and bias, adding new fuel to the push for regulation.

Numerous reports have looked at how advertisers can target ads to exclude certain groups, but this study examines how the ads are delivered once they’re out of advertisers’ hands. Even if an ad is targeted broadly, Facebook will serve it to the audiences most likely to click on it, generalizing from information from their profile and previous behavior. The system builds correlations to find this ideal audience: if techno fans are particularly likely to click on a specific ad for headphones, that ad might be served more to other techno fans in the future, even if it wasn’t an explicit targeting parameter.

"PREVIOUSLY UNKNOWN MECHANISMS” THAT CONTROL WHO SEES AN AD"

The paper (which has not yet been peer-reviewed) is a collaboration between Northeastern University, the University of Southern California, and nonprofit organization Upturn. Its authors tested whether job listings or housing ads with certain keywords or images would
be automatically delivered more often to certain groups, exposing what they call “previously unknown mechanisms” that could violate anti-discrimination rules. The researchers spent over $8,500 on ads that they say reached millions of people, linking to actual job-hunting or real estate sites, among other categories. They ran the same campaigns with different ad copy or photos or at different price rates, checking the demographic breakdowns provided by Facebook on each campaign.

Some simple changes turned up dramatic splits. Housing ads with a photograph of a white family, for instance, were apparently served to more white users than the same ad with a black family. (Facebook doesn’t offer analytics directly based on race, so the researchers aimed ads at locations with different racial breakdowns as a proxy.) An ad for lumber industry jobs was shown to an audience that was 90 percent male, while ads for supermarket cashiers reached an 85 percent female audience. And unlike the ads in a well-known ProPublica exposé, these weren’t specifically aimed at men or women. The only difference was in the text and photos.

Spending rates also seemingly affected who saw the ad. Facebook ads are placed through a bidding process, so a campaign backed by more money may end up reaching more “valuable” users. In this case, an ad with a very cheap campaign had an audience that was 55 percent male, compared to a high-budget campaign, whose audience was over 55 percent female.

HUD’s recent lawsuit claimed that by serving ads based on “relevance,” Facebook is likely reinforcing social inequalities: if most home buyers in an area are white, for instance, Facebook might only show ads to white users. It was presented as an untested theory, but this research offers significant support to the idea.

""WE CAN’T SAY EXACTLY HOW THESE CALCULATIONS ARE DONE."

The researchers stress that they still don’t really know why Facebook’s algorithm is making any of these decisions. “We were able to say with confidence from this study that the content of the ad itself matters a lot to the kinds of people that see it. But we can’t say exactly how those calculations are done,” says Aaron Rieke of Upturn.

Reached for comment, Facebook stressed that it was trying to eradicate bias. “We stand against discrimination in any form. We’ve announced important changes to our ad targeting tools and know that this is only a first step. We’ve been looking at our ad delivery system and have engaged industry leaders, academics, and civil rights experts on this very topic — and we’re exploring more changes,” said spokesperson Joe Osborne.

Osborne said that Facebook was actively studying its algorithms, and he noted that Facebook had supported a US House of Representatives resolution on ethical AI development. He also pointed to Facebook’s earlier ad-targeting changes, which include removing categories that ad buyers could use to discriminate as well as building a tool for users to check all housing ads in its system, regardless of what they see in their news feeds.
This study suggests that changing ad-targeting options might not make these listings meaningfully neutral, and Rieke says that a separate ad database wouldn't go far enough. “It’s certainly a good thing that eventually people will be able to go search all the housing advertisements,” he says. “Even so, I think it matters who Facebook chooses to really push the opportunities in front of.”

"OTHER AD NETWORKS MIGHT FACE THE SAME ISSUES"

Facebook has argued that Section 230 of the Communications Decency Act shields it from liability for advertising content. But one of the researchers' major arguments is that Facebook is single-handedly defining these audiences, and advertisers may have little say over how it’s done. “We didn’t say ‘masculine lumberjacks wanted,’” says Rieke. “We took pains to be very clear and neutral in the language of our test advertisements, and we saw these results nonetheless. This is not an issue where advertisers just need to be more careful about the content of their ads.”

So how would Facebook create a system that could avoid legal scrutiny? It could suspend targeted advertising on posts for jobs or housing, or it could change its targeting system to actively counter bias. It could also shunt these listings to a separate system, like the housing ad database Facebook has promised to build.

For now, we don’t know if this paper will affect HUD’s lawsuit against Facebook; the agency declined to comment, citing restrictions on talking about an active legal dispute. But if the case goes to trial, HUD might seek internal data that would back up the paper’s conclusions.

If a court rules that Facebook’s ad placement algorithm is discriminatory, advertising networks across the web might have to change their practices. The researchers say Facebook’s “walled garden” made it particularly suited for this experiment, but it’s plausible that Google or any other ad platform could display the same biases. “We did not yet measure other advertisers,” says co-author Piotr Sapiezynski. “But we do suspect platforms that try to reach whatever they define as ‘relevant’ audiences might run into this situation.”
What happens next in the housing discrimination case against Facebook?

By Adi Robertson@thedextriarchy  Apr 2, 2019, 1:09pm EDT

Last week, the US Department of Housing and Urban Development sued Facebook for allegedly making housing discrimination easy. It claimed that Facebook violated the Fair Housing Act by letting ad buyers target audiences that included or excluded certain races, religions, or genders. The move took many people by surprise, apparently including Facebook. Soon, it could test how long-standing rules against housing discrimination intersect with the sometimes controversial laws covering web platforms and how they apply to the vast, often little-understood advertising networks that help power the internet.

Facebook has skirted around these issues before. The company had just reached a settlement in a related lawsuit, agreeing to eliminate several targeting categories. It was also working with HUD to address housing discrimination concerns until talks supposedly broke down over a request for user data. Now, the company will be fighting charges that can result in substantial financial penalties and meaningful changes to the way its ad system operates. If HUD establishes that Facebook violated the Fair Housing Act, it’s setting the stage for lawsuits against Google and Twitter, both of which are reportedly under scrutiny as well.

This lawsuit is unusual for a variety of reasons, including the fact that HUD often leaves lawsuits up to local agencies, says Bleakley Platt & Schmidt attorney James Glatthaar, who has dealt extensively with housing discrimination cases. (Disclosure: a Vox Media video team member’s father also works at Bleakley Platt & Schmidt.)
“CRAIGSLIST AND OTHER SITES HAVE BEEN ACCUSED OF DISCRIMINATION”

Glatthaar says HUD may have taken the investigation on because Facebook is a relatively unique case. “Something like this is a national practice, whereas most housing discrimination is fairly local,” he says. “The federal government is one of the few entities that can handle something of this scope.” The agency might also want to specifically address a gray area in the law, make a public statement about housing discrimination, or simply make an example of Facebook.

HUD’s suit isn’t the first case involving an online service. Rental service Roommates.com and classified site Craigslist both went to court in 2008 to fight charges that they’d enabled racist or otherwise exclusionary listings. Both companies claimed protection under Section 230 of the Communications Decency Act, which shields web platforms from liability for user posts. Craigslist successfully made the defense, but Roommates.com was held partially liable because it had offered a survey that included discriminatory questions.

When civil rights groups sued Facebook over housing discrimination last year, Facebook used Section 230 as a defense, saying it had only provided a set of general advertising tools. That argument wasn’t tested in court, however, so we don’t know how courts will treat ad-targeting options like Facebook’s now-defunct “ethnic affinity” selector.

“HUD IS TARGETING ALGORITHMS, NOT JUST CHECKBOXES”

If Facebook defends itself with Section 230 again, this could make for a major test of the rule, which has grown politically controversial in recent years. Over the past year, two court cases — one against Grindr, the other against Yelp — have affirmed Section 230’s safe harbor protections, and a ruling here could help establish just how far they reach. Also, support for the rule among lawmakers is weakening, and a case involving a powerful company like Facebook could spark further debate over it.

HUD is also making some additional claims that could complicate Facebook’s defense. In addition to calling out tools that let advertisers select audience categories, it’s condemning the invisible process Facebook uses to serve ads. “[Facebook’s] ad delivery system prevents advertisers who want to reach a broad audience of users from doing so,” it says, because it’s likely to steer away from “users whom the system determines are unlikely to engage with the ad, even if the advertiser explicitly wants to reach those users."

HUD doesn’t have to establish that these targeting algorithms are designed to avoid showing ads to certain protected classes. It just has to demonstrate that the system effectively makes housing less accessible to these people — a concept known as disparate impact. “If there is an algorithm that just happens to discriminate against racial minorities or gender minorities or whatever, I think it would still be problematic,” says Glatthaar’s colleague Adam Rodriguez. He compares the move to a zoning restriction.
whose text and intent is race-neutral but that directly results in fewer black residents, which would likely still be considered discriminatory.

"FACEBOOK HAS ARGUED THAT FAKE ADS DON’T SHOW REAL HARM"

In its defense last year, Facebook claimed that there were no concrete cases of advertisers denying access to real housing options. Groups like the National Fair Housing Alliance and ProPublica had purchased fake ads to prove discrimination was possible, and Facebook later acknowledged that these groups raised “valid concerns” about its practices. But it argued that their evidence only showed “the possibility that some unidentified third parties may use those tools to place real discriminatory ads.”

Rodriguez says Facebook may have felt confident in refusing HUD’s request for data because it believed the agency’s legal case was weak. But if the case goes to trial, the stakes are still high. HUD is requesting unspecified financial damages, plus civil penalties that can reach $50,000 per violation, and it’s not clear how many violations a court might say Facebook accrued, given its massive size. Also, as the trial progresses, Facebook might have to let HUD comb through emails or other documents, potentially producing hard evidence of people being denied access to real housing listings — or just giving HUD (or other critics) access to potentially unflattering information.

For now, Facebook has said it will “continue working with civil rights experts” on improving its advertising practices. It may keep looking for a settlement with HUD, sending this case down the same path as its last housing lawsuit and leaving the big legal questions up in the air. “I don’t think we can predict how this one is going to turn out,” says Glatthaar. “There’s no obvious answer.”
The Legislation That Targets the Racist Impacts of Tech

A proposed law would make big companies determine whether their algorithms discriminate, but it’s lacking in some big ways.

By Margot E. Kaminski and Andrew D. Selbst
Ms. Kaminski is a law professor and Mr. Selbst is a postdoctoral scholar.

May 7, 2019

In the wake of recent revelations about biased algorithms, congressional Democrats have proposed a bill that would require large companies to determine whether the algorithms they’re using result in discrimination, and work to correct them if they do.

The bill, called the Algorithmic Accountability Act and introduced last month by Senator Ron Wyden, Senator Cory Booker and Representative Yvette D. Clarke, is a good start, but it may not be robust enough to hold tech companies accountable.

When creating a machine-learning algorithm, designers have to make many choices: what data to train it on, what specific questions to ask, how to use predictions that the algorithm produces. These choices leave room for discrimination, particularly against people who have been discriminated against in the past. For example, training an algorithm to select potential medical students on a data set that reflects longtime biases against women and people of color may make these groups less likely to be admitted. In computing, the phrase “garbage in, garbage out” describes how poor-quality input leads to poor-quality output. In this case we might say, “White male doctors in, white male doctors out.”

In March, the Department of Housing and Urban Development sued Facebook for violating the Fair Housing Act. It accused Facebook of keeping some users from seeing housing ads based on machine learning algorithms’ inferences about those users’ race and other characteristics.

Because algorithmic models are incredibly complex, it can be hard for even their designers to know exactly how outcomes come about. But it’s possible to study their design and anticipate their results.

The proposed bill would be a significant step forward toward ensuring that algorithms are fair and nondiscriminatory. It requires certain businesses that use “high-risk automated decision systems” (such as those that predict a person’s work performance, financial situation, or health) to conduct algorithmic impact assessments. This means they must, as Mr. Booker put it, “regularly evaluate their tools for accuracy, fairness, bias and discrimination.”

[Technology has made our lives easier. But it also means that your data is no longer your own. We’ll examine who is hoarding your information — and give you a guide for what you can do about it. Sign up for our limited-run newsletter.]

The bill is a meaningful first step in addressing the problems with algorithmic decision-making. Companies must be pushed to consider and document what goes into algorithm design. They should be pushed, too, to come up with solutions. But the bill is lacking in three main areas.

First is enforcement. The bill largely relies on the agency responsible for consumer protection, the Federal Trade Commission, to issue and enforce regulations about what should be in companies’ impact assessments. The bill grants the F.T.C. new enforcement powers, including the power to require companies to “reasonably address in a timely manner the results of the impact assessments,” but the agency even now too rarely enforces its settlements with repeat privacy violators. Significantly, the bill does not clearly prohibit algorithmic bias or unfairness. It relies instead on the F.T.C. to do so, or on penalties set by existing consumer protection and discrimination law, which may not cover all forms of algorithmic bias.

Second, the bill lacks an avenue for meaningful public input. Technology companies often lack diverse voices and fail to adequately consider social impact, resulting in numerous fiascos — from Google’s image recognition algorithm that classified black people as gorillas, to Amazon’s job-recruiting engine that discriminated against women. In the United States, the typical impact assessment process,
based on environmental law, includes an opportunity for public comments, but this bill would not require them.

The proposal suggests that the impact assessments should be conducted with independent auditors and technology experts “if reasonably possible.” But limiting external input to when it is “reasonably possible” may allow companies to evade public feedback. And discussions about discrimination require input not just from software engineers but from affected communities, legal experts and public interest representatives.

Third, the proposal needs to mandate at least some public transparency for the results of impact assessments. If the results of assessments aren’t public, we can’t learn anything from them. The bill is missing a way for insights to make their way back into broader policy discussions. For example, a common question about algorithmic bias is whether it is worse than human decision-making. Impact assessments can be a tool to help society figure this out — but not if they are kept secret.

There may be good reasons for some limited secrecy; full transparency might be off the table because of concerns about proprietary information or gaming the algorithm. One way to address this failing would be to require the F.T.C. to produce an annual report on the broader lessons it has learned from that year’s impact assessments.

The proposed Algorithmic Accountability Act is a welcome and necessary first step toward governing the secret, often biased algorithms already widely in use across society. But without increased accountability, it may not be effective.

Margot E. Kaminski is an associate professor at the University of Colorado Law School. Andrew D. Selbst is a postdoctoral scholar at the Data and Society Research Institute and a visiting fellow at the Yale Information Society Project.

Follow @privacyproject on Twitter and The New York Times Opinion Section on Facebook and Instagram.
Racial formation, inequality and the political economy of web traffic

Charlton McIlwain

To cite this article: Charlton McIlwain (2017) Racial formation, inequality and the political economy of web traffic, Information, Communication & Society, 20:7, 1073-1089, DOI: 10.1080/1369118X.2016.1206137

To link to this article: https://doi.org/10.1080/1369118X.2016.1206137

Published online: 11 Jul 2016.

Article views: 1685

View related articles

View Crossmark data

Citing articles: 5 View citing articles
Racial formation, inequality and the political economy of web traffic

Charlton McIlwain

Department of Media, Culture, and Communication, New York University, New York, NY, USA

ABSTRACT
Few studies attempt to demonstrate whether and how systemic racial inequality might form on the web. I use racial formation theory to conceptualize how race is represented, and systematically reproduced on the web, and how both may reveal forms of racial inequality. Using an original dataset and network graph, I document the architecture of web traffic, and the actual traffic patterns among and between race-based websites. Results demonstrate that web producers create hyperlink networks that steer audiences to websites without respect to racial or nonracial content. However, user navigation reflects a racially segregated traffic pattern; users navigate to racialized versus nonracialized websites (and vice versa) more than what would be expected by chance. These results, along with disparities in website traffic rankings, provide evidence of, and demonstrates how a race-based hierarchy might systematically emerge on the web in ways that exemplify disparate forms of value, influence and power that exist within the web environment.

ARTICLE HISTORY
Received 28 December 2015
Accepted 22 June 2016

KEYWORDS
Internet; racial segregation; inequality; social networks; algorithms; racial formation

Conversations about racial inequality on the web are longstanding. However, since the mid-1990s, these conversations proceeded along disparate paths, without a central connecting thread, theoretical model, or substantial evidence to suggest whether and how racial inequality gets produced on the web. Three such conversations stand out. First, differential access to the Internet in the mid-1990s sparked conversation about a racialized digital divide. (Hoffman & Novak, 1998). Despite narrowed access gaps, access questions related to differential digital knowledge, preparation and participation persist (Hargittai, 2010; Schradie, 2012).

A second early conversation focused on problematic racial formations such as identity tourism (Nakamura, 2002), White’s domination of racial discourse in online spaces (Burkhalter, 1999; Kendall, 1998), digital Othering (Chow-White, 2006) and the proliferation of White Supremacist organizations online (Daniels, 2009; Futrell & Simi, 2004). Other work focused on various diasporic communities forming online that used Internet affordances to connect with other users of color, and to control racial group representations in

CONTACT Charlton McIlwain cdm1@nyu.edu Department of Media, Culture, and Communication, New York University, East Building, 239 Greene Street, 8th Floor, New York, NY 10003, USA
© 2016 Informa UK Limited, trading as Taylor & Francis Group

A third conversation is least developed, but most significant for our ability to conceptualize and document racial inequality online. Two significant peaks in an otherwise flat scholarly literature mark attempts to address race, technological systems and systematically produced inequality. Friedman and Nissenbaum (1996) proposed that computer systems might propagate a variety of biases, racial and cultural biases included. The authors built a comprehensive framework to identify, assess and potentially remediate such biases; however, the next scholarly work connecting the technical infrastructure of computer systems to racial productions does not arrive until Tara McPherson (2011). She links two narratives that emerged simultaneously, yet independently in the 1960s. One first focused on the development of the UNIX operating system, and the other, post-Civil Rights racial discourse. McPherson argued that the post-civil rights racial logic resembles the lenticular logic underpinning UNIX. Why did it take nearly two decades (between Friedman and Nissenbaum and McPherson’s work) to return to questions about the link between digital technology and the systematic production of race? Why did it take so long to imagine a specific relationship between the two in ways that might direct our thinking about how digital technology might shape not only racial identity or racial discourse, but also potentially reproduce racial inequality?

I outline in the following pages a theoretical foundation for exploring how race is represented, and systematically reproduced on the web, and how both may reveal forms of racial inequality. First, I argue that we must conceptualize racial inequality online as a racial – more than purely technological – formation. Second, I use a spatial analytic to briefly outline the dominant political and economic logic of web traffic that underpins today’s web. Third, I draw on an original dataset and network graph to document the architecture of web traffic, and the actual traffic patterns among and between race-based websites. These data demonstrate the ways in which a race-based hierarchy might systematically emerge on the web in ways that help us to identify disparate forms of value, influence and power that exist within the web environment.

**Racial formation and web inequality**

I attempt to develop a framework for conceptualizing and demonstrating how racial inequality might be produced and systematically propagated on the web. The myriad ways that people use online platforms and tools to accomplish race-based or racist goals is not my concern here. Rather, I aim to understand how racial advantage and disadvantage might get systematically produced within the web’s structural edifice – one built to assist users in navigating the web and shuttling groups of users in and out of various web spaces.

Recent attention to algorithmic power (Beer, 2009; Gillespie, 2014), the values encoded in them, and the outcomes they produce, is central to identifying how racial inequality might get produced online. However, I argue that racial formation theory provides a broader, and more appropriate framework for understanding how race works on the web. To be sure, the algorithmic lies at the heart of what I think is most significant
here. But algorithms are agnostic tools, used to engineer solutions for which there are antecedent problems or interests – both of which emerge from a set of historical contexts and circumstances. To fully understand how a technological system such as the Internet might produce tangible forms of race-based inequality, we must consider how the Internet developed as a part of a longstanding history and process of racial formation – the complex, racialized historical contexts, circumstances, interests and problems that predate, but may either be exacerbated or corrected by the web’s technological environment.

**Technological formation as racial formation**

Racial formation reflects the ways that individuals and communities conceptualize, represent and articulate the meaning of race, and its significance. Racial formation also encompasses the means by which racial meanings are systemically incorporated, circulated and appropriated by institutions that automatically produce and reproduce, sustain or challenge an existing social and political order that confers access, privilege and power along racial lines. Racial formation theory (Omi & Winant, 1994) helps us to understand the existence, development and persistence of racial inequality by detailing the ways that representational systems (Hall, 1997) and societal institutions, draw on prevailing racial common sense to produce systematic differential advantages and disadvantages that intentionally or unintentionally flow to racial group members.

Identifying a guiding logic animating both racial and technological systems is a useful starting point to understand the Internet as a technological formation that is part of a broader historical trajectory of racial formation. Broadly speaking, a ‘logic’ is a coherent system of principled reasoning leading to specific conclusions. In computing, logics function to engineer how machines operate. To suggest that both racial formations and technological formations share a central logic is to say that each are, by definition, technological. Technologies – whether we’re referring to race or computers – are a means to an end, a tool to produce specific outcomes.

McPherson (2011) argues that a lenticular logic which privileges concealment and modularity animated narratives that birthed both personal computing, and prevailing post-civil rights narratives emerging in the mid-sixties. Lenticular logics mask core operations, be they computational or political. ‘UNIX’s intense modularity and information-hiding capacity were reinforced by its design, that is, in the ways in which it segregated the kernel from the shell,’ Mcpherson writes, adding that ‘the second half [of the twentieth century] increasingly hides its racial ‘kernel,’ burying it below a shell of neoliberal pluralism’ (Mcpherson, 2011, p. 29).

My point here is twofold. First, while the lenticular was central to computing applications, the logic itself was already socially established. Two significant racial projects were already underway at the time of the computing debates that McPherson cites. First, there was a normative shift toward racial equality. Mendelberg (2001) argues that this shift did not necessarily signal widespread racial attitude change so much as it rendered explicit racial prejudice and discrimination less publicly tolerable. The transformation did not diminish the technological power of racialization. It simply necessitated effective means to conceal the ways that race animated institutional systems.

The second racial project manifesting this lenticular logic is the neoconservative racial project (Hall, 2005), which rearticulated landmark civil rights legislation in the 1960s as
the culmination – rather than the beginning or continuation – of the civil rights struggle. This reframed narrative accomplished two goals: it truncated the movement and singled out legislative action as its primary form of political action. This narrative covertly erases the reality that the civil rights struggle persisted since Reconstruction – what Hall and others refer to as the long civil rights movement; obliterates the complex and integral connection between racial struggles, labor struggles, and the freedom struggles of other racial groups and women; minimizes racial inequality’s institutional breadth that extends to voting, housing, education, criminal justice, employment, banking, media and beyond; and finally, helps to establish the ideology of colorblindness as the new racial common sense. This extended far beyond simply providing the veneer of neoliberal pluralism; the narrative made it possible to steadily erode the need, desire and ability for government to provide public policy interventions to mitigate race-based inequalities. As such, the neoconservative racial project normalizes racial silence, such that racial discourse in policy and political contexts is seen as aberrant and deleterious.

I reiterate, but also extend McPherson’s central argument. The lenticular logic that pervades both computing and the sociopolitical field in the U.S. is not merely coincidental. The concrete neoconservative racial project that elevates colorblindness and negates public motivation to identify or recognize racial inequality and disparities, and intervene to mitigate it, extends specifically and directly to today’s digital technologies, particularly the online environment. Understanding whether and how racial inequality might emerge in the contemporary online environments necessitated viewing the development and contemporary workings of the Internet as part and parcel of this neoconservative racial project.

**Racial formation, spatial logics and the political economy of web traffic**

The neoconservative racial project extends to the web, which – in addition to the lenticular – mobilizes a spatial logic online that mirrors the spatial/geographical production of racial inequality offline. Historically, space has been the tool by which racialization, and the dominant system of White Supremacy has been both created and sustained. White elites, and White commoners who benefited from whiteness, configured and used space to transform the fantasy of racial distinction into a meaningful reality. By structuring the spatial relationships between people and valuable resources (land, property, infrastructure and institutions), geography formed the foundation for persistent racial problems: segregation, urbanization, ghettoization, race migration, racial zoning, redlining, blockbusting, bussing, integration, gentrification, steering, property tax funding of education, gerrymandering, racially exclusive social networks and the like. Racial inequality was created and has been maintained by defining, assigning value to, and strictly policing inclusion and exclusion into and from racially marked spaces that confer power, privilege and access (Briggs, 2005).

Race and space have been so intimately intertwined in the U.S. that they are essentially co-constitutive (Delaney, 2002; Lipsitz, 2007). Space makes race, and race – its meaning in both concept and everyday life – expresses itself in particular kinds of spaces and spatial configurations. As Lipsitz highlights through what he dubs a White vs. Black spatial imaginary, the spatial alignment with race has resulted in two broad outcomes. First, space is racially marked for the purposes of identity and identification, in ways that simultaneously
serve both maintaining white supremacy and building racial solidarity among people of color. Second, the racial marking of space is used to create networks, consisting of the geographical arrangement of people, power and resources. These spatialized networks systematically control the advantages gained by being granted access to certain networks, and disadvantage through systematic exclusion or unequal access.

I focus on these racial-spatial relationships for two reasons. First, both the web’s nomenclature and interface is largely spatial (Graham, 2013; Graham & Zook, 2011). Second, if space structures the web environment in ways that it structures social, political, racial life offline, then it makes sense to ask: how do configurations of web space produce racial meaning? Furthermore, how might these configurations systematically produce race-based inequality that is specifically tied to the web’s spatial and economic structure?

**Racializing ‘sites’**

Scholars have long critiqued the web’s presumed racelessness as an assumed Whiteness (Burkhalter, 1999; Kendall, 1998; Nakamura, 1995), and have demonstrated how racial representations align with longstanding racial stereotypes (Daniels, 2009, 2013). But such anecdotal examples can still be dismissed as the work of individual web users acting out personal prejudices in online spaces. If we want to know how race and power, advantage, disadvantage and inequality are systematically embedded in the web’s operating structure; however, we must first determine how race itself is systematically – rather than anecdotally – represented online. We must ask, what significance and use-value does race have in the political economy of the web, and how does racial meaning and value get systematically produced and circulated throughout the web?

There is a distinction to be made between the social, political and economic structure of the web, as opposed to the structures that have historically governed everyday social life. Historically, racial formations presume the existence of both individual human beings and racial group categories. Individuals exist by virtue of our corporeal being. The individual also exists through legal inscription and political authority that grants political subjects the rights and responsibilities afforded citizens of geopolitical entities. Individuals vote, own property and are held accountable for legal violations, for example. Race, and racial categories historically exist much the same way. Rights, power, privilege, advantage, access and disadvantage: all – more or less – are social characteristics afforded raced individuals and is accounted for through bureaucratic systems that have an interest in tracking both (Morning, 2011; Morning & Sabbagh, 2005). I labor the point to highlight the contrasting reality that, for all intents and purposes, neither the individual, nor racial categories are similarly salient within the basic structure of the web.

Understanding how race systematically figures into the structure of the web begins with the ‘site,’ not with the individual, though there is a bit of a paradox here. In one respect, sites that thrive on individual users as the source of content used to produce audiences seem averse to identifying, or even allowing users to systematically identify themselves based on race and/or ethnicity. Web users know that trafficking the web almost always includes opportunities or requirements to reveal personal information. In either case, rarely, if ever are users asked to identify themselves by race/ethnicity. The widespread lack of opportunities to identify by race/ethnicity reveals the unstated presumption that race either does not exist and/or is not useful. To the degree that it is seen as both real
and useful – the lack of opportunity to identify may reveal a belief that such information should not be used to structure the web environment.

Sites replace individuals in the web’s central organizing framework; neither you nor I is of interest to those who traffic in it, except in the aggregate. The web’s central function is to facilitate the movement of audiences through a network of sites where a secondary commodity gets produced – content (Van Couvering, 2011). Content attracts audiences, but sites are what are most visible on the web. Users are counted and accounted for only as members of an aggregated audience being shuttled around from one source of content to another. As such, two demographics are principally measured in the political economy of the web – audience traffic, and the ranking of the individual sites that are the sources and destinations for audience traffic.

How then does racial identity and identification operate at the level of the site? How do we know whether a site is racialized, and what racial group it is associated with? What are the means by which sites create racial identity and/or are identified as being racially associated with some group or another? Most importantly, how does racial identity and identification at the level of the site get systematically (and automatically) produced and circulated?

Two examples of how racialization functions in everyday life (offline and on) serve as a point of comparison here. First, in societies where race is salient and accounted for, the process of identity and identification is, more or less, reciprocal. Bureaucratic agencies require citizens to identify. The agency sets the categories, citizens choose from the given options and the bureaucracy collects, counts, aggregates and uses the data for various political purposes. Aside from bureaucratic forms of racialization, spaces become racialized based on the flow of bodies in, out and occupying particular spaces. Black spaces are created when a critical mass of black people, for instance, occupy a space. Their occupation, and their codification as ‘black space’ establish a particular real or assumed character onto a space. It also brings with it a certain valuation (or devaluation). This process of creating racial space with particular cultural associations and valuations underlies things such as neighborhood segregation, gentrification, steering (Briggs, 2005).

A second example demonstrating how websites sites become racially marked in systematic ways can be encapsulated in the website categorization project conducted at DMOZ.org. Categorizing both sites and content are central to how the web works, particularly for search engines, which dominantly mediate web traffic. However, only one attempt has been made to systematically account for, and categorize sites according to racial/ethnic identity. DMOZ.org has, since 1999, relied on a vast team of editors to categorize the web’s vast array of websites. The DMOZ project comes closest to approximating the way that racial identity and identification has worked historically, in the more bureaucratic sense. On the one hand, DMOZ editors have the power to create the categories and provide instructions to other editors about how to determine what fits in the category. On the other hand, it is a site’s content – or information gleaned or inferred from the site – that is the primary basis for categorization. Sites (more specifically their owners, and/or agents) determine content, not DMOZ. To be sure, racial and ethnic categories are not primary categories in DMOZ’s system. One can arrive at varying race and ethnicity site categories from several different starting points in DMOZ’s path structure. The most direct route to the greatest number of sites classified by race/ethnicity is: Society > Ethnicity, which leads to several categories of mostly white, European-based ‘ethnic’ group categories such as
Albanian, French and Romani. One can also get to Hispanic/Latino and African-American on this path: Society > Ethnicity > African > African-American, where one finds 308 websites. A brief description of the sites categorized here states that,

Websites maintained by African Americans, or focusing on African Americans. African Americans are generally people who are classified as citizens of the United States, but also have aboriginal African ancestry. Sites about people born in Africa but living in the United States may also be listed here.\(^1\)

This example is instructive in several ways. First, while editors are given some freedom to categorize within certain parameters, there are some clear characterizations and valuations being imposed in describing/defining the category’s boundaries. One can readily see how such a description might be problematic or controversial. Second, the description provides too much flexibility to be of much reliable use. Is a site owned and maintained by African-Americans an African-American site, even if its content does not focus on African-Americans? What about U.S. Citizens who have some ‘aboriginal African Ancestry,’ but would not otherwise classify themselves as African-Americans? Are native Africans living in America African-American? White South Africans living in the U. S.? U.S. citizens with Algerian roots?

These highlight just a few limitations of DMOZ’s model for racially categorizing sites. But they are enough to see why it is a failed project in terms of providing a systematic way to account for race within the web’s structure. It also demonstrates, perhaps, why various constituents of the web appear agnostic about, or even hostile to such a project altogether. Despite this, and even though DMOZ may not the best way to account for race on the web, it is significant in one other particular and novel respect. It demonstrates how race could be systematically accounted for, and relied on, within the web’s site-content-traffic structure. The metrics used to rank sites and measure traffic to, from and between them rely on sites whose content is identified and categorized. At least one (and likely more) of the prominent Internet traffic measurement enterprises – Alexa.com – is built on top of DMOZ’s categorical architecture. To understand how a systematic reproduction of sites categorized by race might work, take, for example, the popular African-American focused news site, theroot.com. Alexa ranks the site at 2474.\(^2\) This ranking includes in its assessment the category to which the site belongs, as traffic is measured in comparison to similarly categorized sites. Furthermore, Alexa uses the top-level category as the primary basis for comparison. In DMOZ’s category scheme, theroot.com – despite the fact that its content is clearly targeted toward African-Americans and Black-themed issues – is not categorized as a race-based site. Its category tree structure is news > magazines and e-zines > theroot.com. Thus, the site’s 2474 ranking is based on sites within the news category.

Two consequences of this layered classification/ranking scheme are worth pointing out. First, a site’s traffic rank is related to the category to which it belongs, and a site’s prominence within a ranked search engine is a result of both its traffic rank, and the sites in the category to which it compares. If one searches the keyword ‘news’ in Google, for instance, it may come as no surprise that theroot.com does not rank in the total number of 368 sites returned in its results. In this case, a site like theroot.com is rendered virtually invisible and ghettoized when searched according to its identified category (as a nonracial, news site). The second thing to point out is that somewhere in the web’s systemic architecture, race and blackness is being associated, and attributed to the site. How do we know? If
you alter the search just slightly using the phrase ‘black’ or ‘African-American’ news (for which there is no category), then theroot.com falls somewhere within the top three sites returned in Google’s search results.

So why does theroot.com rank at the top of a search for black news when no such category exists, and does not rank at all within the category to which it has been assigned? One likely – though not definitive – answer is that the categorization process is taking place through a different and less transparent process than with DMOZ and Alexa. For one, Google draws on a different source for characterizing and categorizing site content. DMOZ, and by extension, Alexa, relies on human categorization into discreet (and contradictory, and messy) categories. A search engine such as Google relies on coded contextual information – rather than human beings – to render a site’s identity. This code – referred to as meta-tags, are pieces of html code that site producers write into their webpages. These meta-tags – such as title tags, description tags, image tags, snippet tags, and the like – can communicate many things to Google. Such tags give site producers the power to brand sites using descriptive text and imagery they feel best communicates the site’s identity.

The second part of the explanation for why and how the site finds its way to the top of the Google search results for ‘black news,’ rather than just ‘news’ is that Google categorizes sites automatically, not only by collecting and associating content from a page’s various tags, but by associating tagged content with those found on other websites, and by associating this content with search queries. While we do not know how the algorithms specifically operationalize this content to arrive at a particular result, Google’s instructions about how to maximize visibility within its search results provide a clear enough indication that these and other forms of metadata are what it uses to both categorize and rank the web pages it indexes.  

So what does all of this tell us about racial identity and identification on the web? First, the relationship between website categories, traffic rank and search results/rank demonstrate how information about race/ethnicity can be systematically passed along (or withheld) in the web environment, in ways that approximate historical methods of doing so. Second, the inconsistencies between racial or nonracial content categories, and the ability to search for and find catalogued sites based on the presence of racialized content demonstrates a kind of ambivalence about race on the web. Race may not be a significant variable for cataloguing sites, but at the same time, the ability to locate race-based content is clearly valued. This is to say that one feature of the online environment cloaks the significance of race in a veil of supposed colorblindness. Race is apparent, but not transparent.

Finally, my brief discussion thus far about web traffic, search engine rankings and the variety of metrics that figure into those calculations demonstrate the ways that racial identity and identification on the web may have more significant implications beyond personal or collective preferences about representation.

**Racial inequality and the political economy of the web**

The complicated, confusing, contradictory and haphazard manner in which racial identity and identification functions on the web necessitates a specific racial project. To understand whether and how racial inequality might play out on the web, then we must construe race in particular ways to even render answering the question possible. To identify racial
inequality online requires: a strategic way to define and systematically identify sites by race; identify how audience traffic flows within and between racial and nonracial sites; means to measure traffic between varying racial site categories and identify disparities between them; identify what accounts for the traffic disparities; and, finally, describe what implications follow from these disparities. In sum, we must identify and determine how racial disparities in site traffic might significantly impact the primary constituents of the web – both site ‘owners’ and site audiences in specific material and, perhaps, nonmaterial ways. In the remainder of this paper, I focus primarily on the first two items above, and, to a limited degree, the third.

**Race and traffic flow**

Consistent with Wu and Ackland (2014), website networks can be conceptualized in two ways: navigational (hyperlink) networks or clickstream networks. Navigational networks consist of outlinks and inlinks to websites and seek to direct user’s navigation patterns within and between websites, based on web producers’ strategic motivations. Clickstream networks, however, consist of users’ actual navigation patterns. When we consider these types of networks in the context of race, the critical question is this: do producers or traffickers of race-based websites seek to create, and/or actually navigate among racially homophilous networks? Hopmophily pervades social life across multiple contexts, both offline (McPherson, Smith-Lovin, & Cook, 2001), and online (Appiah, 2004; Nowak & Rauh, 2005; Thelwall, 2009), making it reasonable to suspect that both the directed networks created by hyperlinks and the clickstream patterns of actual users would reflect this tendency toward segregation that continues to characterize contemporary residential patterns in the U.S.

I generated two networks to test this hypothesis. I used VOSON, and beginning with Alexa’s list of ‘Top Black Sites’ as seed sites, I produced a network based on inlinks and outlinks to each. Fifty-six initial seed sites produced a network consisting of 5596 nodes (sites). As a data quality control measure, I reduced the number of nodes significantly by filtering out those with fewer than two connections to other nodes in the network. What remained was a network consisting of 961 nodes, with 3034 links between them.

Beginning with the same seed sites, I also developed a clickstream network using data from Similarweb, a web analytics service that provides Internet traffic data that track both the amount of traffic websites generate, as well as the directional flow of traffic from and to websites. The clickstream network consists of the seed sites, the sites people arrive at the seed sites from, and sites where people go after visiting the seed sites. The total network consisted of 12,524 Nodes and 25,029 edges. Again, for ease of analysis, I reduced the size of the network into a sample consisting of nodes with two degrees or more. The resulting network consisted of 3233 nodes, and 15,884 links between them.

Given my general hypothesis, the linking structure in these networks may reflect the desire to steer traffic, or navigate in a homophilous/inbreeding pattern based on one of three characteristics: race, content category or traffic rank. The first would reflect motivations to affiliate based on racial similarities characteristic of site owners and/or content, the second solely on content similarity, and the third based on what we might call capital gains – the ability to increase reputational capital (search visibility) by affiliating with others high value sites. To test the above hypothesis, I constructed three separate variables.
Race is a binary variable constructed by assessing a combination of primarily linguistic factors that identified whether race related terminology was found either in the title, description or keyword meta-tags. Sites where race was found in either of those data categories were categorized as racial, as opposed to nonracial. Content category variables were constructed as a binary variable corresponding to whether the site either belonged or did not belong to the following highest represented site categories: arts and entertainment, news media, people and society, Internet/telecomm or business and industry. Finally, I collected data on each site’s traffic rank (MozRank), which assesses website prominence based on a number of factors.

I used UCINET, to calculate homophily. The program uses a spatial, autocorrelation algorithm that compares observed versus actual links found among and between racial and nonracial sites. That is, it calculates what we might call an opportunity structure (expected number of connections within and between groups based on chance) and then compares whether the actual connections significantly exceed or fall below what was expected based on chance. A fully homophilous linking pattern would look like this: each group would have a significantly greater number of connections within the group than what would be expected by chance, and the links between groups would be significantly fewer than what would be expected by chance. This is what I designate as a segregated traffic pattern. The networks could exhibit two other patterns. One is an integrative pattern, where nonracial and racial sites have significantly fewer links within, and significantly greater number of links between them. The third potential pattern is what I designate as a status quo traffic pattern. In this scenario, sites do not link in any significant way beyond what would be expected by chance.

**Racial homophily**

Using the above criteria, the hyperlink network is not racially homophilous. Seen in Table 1, links between nonracial sites in the network are fewer than what would be expected by chance. These sites connect more frequently to racial sites, though not in a significantly greater way. Differently than the nonracial sites, racial sites connect more frequently to other racial sites. But again, the degree to which they connect with other racial sites is not significantly greater than what would be expected by chance. Thus, the hyperlink network exhibits the status quo linking pattern. That is, content producers on these sites do not go out of their way to steer audience traffic to other sites based on the racial or nonracial nature of the site. In a network consisting of racial and nonracial sites, links form randomly between them with little effort to intervene. Again, this is a hyperlink network, meaning that the links function to direct users’ navigation to other sites. The presence of a link does not guarantee that a user will actually click on a link and follow it to the specified site.

While content producers of the sites in this network do not direct traffic flow to other sites on the basis of race, the opposite pattern is apparent in the clickstream network that

<table>
<thead>
<tr>
<th>Table 1. Navigation network racial homophily.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation network</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Nonracial/nonracial</td>
</tr>
<tr>
<td>Nonracial/racial</td>
</tr>
<tr>
<td>Racial/racial</td>
</tr>
</tbody>
</table>
indicates the incoming and outgoing flow of actual users to and from sites in the network. This network exhibits a segregated pattern. Seen in Table 2, users visiting nonracial sites visit other nonracial sites with greater frequency than what would be expected by chance, and visit racial sites less frequently than what would be expected by chance. Visitors to racial sites, more than what would be randomly expected, visit other racial sites. In each case, visitors to racial and nonracial sites visit similar websites in ways that exceed the opportunity structure.

**Content homophily**

Given that users navigate to racial and nonracial sites in homophilous ways, it is important to determine whether such patterns exist when other reasons for group formation exist. While users might navigate the web within and across racial boundaries, content preferences also motivate navigational patterns. Content categories in my analysis include: arts and entertainment, people and society, Internet/telecommunication, and business and industry.

From the standpoint of the hyperlink network, the question is whether web producers direct web traffic in ways that steer users toward sites with similar or different categories of content? The data in Table 3 show mixed results. Business and Internet categories reflect status quo patterns, where connections with and between sites in each category do not depart from random expectations. Arts and entertainment sites reflect an integrative pattern, where links to sites outside the category prevail. However, both the news media and people and society categories reflect much more of a segregated pattern, where within-group linking patterns prevail. The homophilic tendencies in these two categories might reflect the broader patterns of polarization found in news and political information consumption patterns (Adamic & Glance, 2005; Lawrence, Sides & Farrell, 2010). While I only test five of many content categories, the results here suggest that web producers, by and large, freely direct users across content categories. Two particular categories, however, reflect the opposite pattern.

The clickstream network reflects no ambivalence. The results in Table 4 are completely uniform; each category reflects a status quo pattern in which users navigate between sites in expected ways, unbound by stronger or weaker connections to sites based on the content category to which it belongs. So, while users do navigate between sites based on their racial characteristics, they, in this instance, do not create such clickstream patterns based on content.

**Ranking homophily**

The final possibility I test here is whether users connect to sites based on their perceived or actual prominence, especially when a search engine is mediating that navigation. In that

<table>
<thead>
<tr>
<th>Table 2. Clickstream network racial homophily.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clickstream network</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Nonracial/nonracial</td>
</tr>
<tr>
<td>Nonracial–racial</td>
</tr>
<tr>
<td>Racial–racial</td>
</tr>
</tbody>
</table>

*Indicates a significant difference between expected and observed frequencies.
case, the sites most visible and most likely to be visited by users are those at the top of search ranking results. If this economy of visibility is prominently at work, we might expect it to be reflected in two opposing patterns among hyperlink and clickstream networks. In the former, a nonhomophilous pattern may be most desirable, reflecting a web producer’s motivation to link to higher ranking sites for the purpose of gaining the capital it might receive from a reciprocal link. In the latter, a segregated traffic pattern would indicate the tendency on the part of users to traffic within neighborhoods of low- or high-ranking sites – creating and maintaining a kind of class-based division between highly and less prominent sites.

Contrary to the possible scenario outlined above for the hyperlink network, the homophily results reflect a status quo pattern, meaning that site producers are not directing users to other sites based on the rank or visibility of the site. Producers link to low- and high-ranking sites in random pattern without purposeful attention to whether linked sites are consistent with the linking site’s own rank.

However, as seen in Table 6, the clickstream network tends toward segregation. While this tendency is present, the in-group connection patterns do not, in each instance, depart

<table>
<thead>
<tr>
<th>Navigation network</th>
<th>Expected</th>
<th>Observed</th>
<th>Difference</th>
<th>P&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not A&amp;E/A&amp;E</td>
<td>1470</td>
<td>470</td>
<td>−1000</td>
<td>.00*</td>
</tr>
<tr>
<td>Not A&amp;E/A&amp; E</td>
<td>1217</td>
<td>1735</td>
<td>518</td>
<td>.00*</td>
</tr>
<tr>
<td>A&amp;E/A&amp;E</td>
<td>250</td>
<td>732</td>
<td>482</td>
<td>.00*</td>
</tr>
<tr>
<td>Not news/not news</td>
<td>2143</td>
<td>2511</td>
<td>368</td>
<td>.00*</td>
</tr>
<tr>
<td>Not news/news</td>
<td>732</td>
<td>426</td>
<td>−305</td>
<td>.00*</td>
</tr>
<tr>
<td>News/news</td>
<td>62</td>
<td>0</td>
<td>−62</td>
<td>.00*</td>
</tr>
<tr>
<td>Not PS/not PS</td>
<td>2547</td>
<td>2756</td>
<td>209</td>
<td>.00*</td>
</tr>
<tr>
<td>Not PS/PS</td>
<td>376</td>
<td>181</td>
<td>−195</td>
<td>.00*</td>
</tr>
<tr>
<td>PS/PS</td>
<td>14</td>
<td>0</td>
<td>−14</td>
<td>.07*</td>
</tr>
<tr>
<td>Not Internet/not Internet</td>
<td>2530</td>
<td>2675</td>
<td>145</td>
<td>.17</td>
</tr>
<tr>
<td>Not Internet/Internet</td>
<td>392</td>
<td>262</td>
<td>−130</td>
<td>.18</td>
</tr>
<tr>
<td>Internet/Internet</td>
<td>15</td>
<td>0</td>
<td>−15</td>
<td>.06</td>
</tr>
<tr>
<td>Not business/not business</td>
<td>2697</td>
<td>2800</td>
<td>103</td>
<td>.20</td>
</tr>
<tr>
<td>Not business/business</td>
<td>235</td>
<td>137</td>
<td>−98</td>
<td>.20</td>
</tr>
<tr>
<td>Business/business</td>
<td>5</td>
<td>0</td>
<td>−5</td>
<td>.24</td>
</tr>
</tbody>
</table>

*Indicates a significant difference between expected and observed frequencies.
significant from chance. I would argue that the one significant departure here – that those navigating low-ranking sites have fewer connections with high-ranking sites – provides at least ample reason to suspect that a kind of segregated, class-based clickstream pattern may exist within the racial web. That is the general segregated pattern exists here, with the qualification that the number of cross-race connections is the only dimension where the observed number of connections differs significantly from chance (Tables 5 and 6).

**Discussion**

The idea that racism and racial inequality may pervade the online environment is not a novel concept. But until recently, evidence supporting this contention has been anecdotal, largely limited in its focus on interpersonal and representational forms of bigotry that now routinely circulate throughout the web. By contrast, a century worth of attention and documentation lays out the ways in which the concept of race originated, was imbued with social and political significance, and the ways that racial meanings have and continue to be systematically produced and reproduced in ways that deliver advantage and disadvantage along racial lines. To the degree that everyday contemporary social life increasingly tethers us to the online environment, I argue that it is behooves us to more fully explicate how race pervades the technological system that is the web – a system that increasingly mediates commerce, politics, education and interpersonal social relations.

Why has such a large scale racial project not been undertaken before now, when we have known from the very beginning that race is a salient feature of life on the web? I would simply argue that our technological formations matured to the point of existing racial formations. When Friedman and Nissenbaum wrote about bias in technological systems in 1996, the commercial Internet was in its relative infancy – so much so that the examples they appealed to revolved around things such airline reservation systems or bureaucratic computer systems used to deliver services or rewards to specific constituencies. It took time for the web to grow into the expansively connected and fully interactive medium with its own economy and a critical mass of users to make it a place where race may be critically significant. Race was apparent on the web as it existed in 1992.
What I attempted to do in this paper is to take a small step toward outlining where and how we might go about trying to understand how race gets systematically incorporated into the web’s environment, and in what way(s) this system might produce advantages and disadvantages across racial lines. The site is central to the web’s economy. It is the place where content resides. It is a primary ‘place’ where race is encoded, apparent and, more or less, visible. Sites and search engines are the two dominant arbiters of web traffic – shuttling, directing, steering or otherwise enabling the movement of people and their data through the web’s vast network. It is the place I argue we should look to investigate whether and how racial inequality might get produced on the web in ways that might comport with historical and geographical racial formations.

The first question I attempted to address was essentially, how do we account for race in this environment. My approach to this question is to look to the principal way that the system provides individuals/entities the opportunity to identify – at the level of the site, using self-described tags to characterize the site and its content in ways that are most consistent with site producer’s own interests and identity. This makes it possible to accomplish one critical component of identifying and measuring racial inequality: racial classification.

The second question I attempted to address is: what is the spatial relationship between these now racialized sites (racialized, to some degree by the sites themselves or by me through this process of categorization). This question is consistent with the historical fact that racial inequality was largely produced and systematically reproduced through spatial relations. The evidence in these data is by no means conclusive, but they suggest a tendency toward racially segregated site navigation. The model evident in these data – with some qualifications – suggests that two things are going on at once in terms of what we might consider the ‘built’ environment of the web that seeks to steer audience traffic, and the actual traffic patterns of users as the move to and from sites around the web. Web producers seem to build pathways providing equitable access to sites, without concern for the racial nature of the site. This might produce truly equitable traffic patterns if users only – or even primarily – relied on site links to direct the flow of traffic. But other things intervene – namely, individuals’ own prerogatives, search engines or a combination of both. The lesson is, just because people build a road to get from point A to B does not mean people will choose to drive on it, or use it to go from point A to C, when C is a destination that comports more with their individual preferences.

In the networks I analyzed here, such diversions most likely occur when the dividing factor is race, more than the content category to which a site belongs. Layered on top of these racially segregated traffic patterns is a tendency toward a class-based segregation as well, where users tend to navigate to sites with the same, rather than differing status. These two characteristics together – segregated traffic patterns divided by race, and the ranking of sites – form a foundation where inequality along racial lines might exist. And it sets the stage for inequality to exist in its truest form, where disparate outcomes along racial lines are coupled with specific advantages or disadvantages. In the data used for my analysis, it is quite easy to demonstrate that disparate outcomes exist with respect to racial and nonracial sites and their rankings: The mean rank of nonracial sites is 4.5, compared to 4.2 for racial sites. Despite what seems like a miniscule numerical difference in mean scores, the difference is statistically significant at the highest level
There is a clear disparity between site rankings based on their racial or nonracial classification here.

So what do we do with these findings? What significance do they have for better understanding how racial inequality might be produced in the online environment? First, I argue that these findings demonstrate that two key variables that have historically contributed to racial inequality are present within the web’s environment: segregated traffic patterns (that lead to segregated destinations); and disparate valuations of those destinations. This finding supports the notion that race may play a significant role in arbitrating systems of value, access an opportunity online in ways that they have historically done so offline. The findings in this paper are significant as well because once we know that segregation and disparate value exist in the online environment, we know what questions remain for research to ask and answer to fully determine whether and how racial inequality may get produced on the web. These remaining necessary, and most significant questions are these: what actual value do site rankings possess? What traffic advantage(s) are to be gained from having higher site rankings? What disadvantages(s) are there to having lower site rankings? Finally, what are the real implications of, or – differently stated – what is the tangible ‘harm’ for a site (and, presumably its human owner(s)) that is disadvantaged in this traffic network?

In the same way that addressing such questions are a necessary condition for legal action under the disparate impact theory of discrimination (Morning & Sabbagh, 2005), demonstrating disparate outcomes based on race within the online environment, and establishing the tangible implications/harms they produce will help us to determine more precisely how racial inequality may continue to be systematically produced within the web’s political economy. Thus, future research should both replicate aspects of this study to determine whether similar patterns of racial segregation and disparate site valuations are present and develop new research designs aimed at determining the tangible gains and harms produced by such patterns of web traffic.

Notes

Disclosure statement
No potential conflict of interest was reported by the author.

Funding
This work was supported by IDEAs Challenge Grant from NYU’s Steinhardt School of Culture, Education and Human Development.
Notes on contributor

Charlton McIlwain is an Associate Professor of Media, Culture, and Communication at New York University. His work focuses on the intersection of race, media and politics [email: cdm1@nyu.edu].

References


Algorithmic decision-making has become synonymous with inexplicable decision-making, but what makes algorithms so difficult to explain? This Article examines what sets machine learning apart from other ways of developing rules for decision-making and the problem these properties pose for explanation. We show that machine learning models can be both inscrutable and nonintuitive and that these are related, but distinct, properties.

Calls for explanation have treated these problems as one and the same, but disentangling the two reveals that they demand very different responses. Dealing with inscrutability requires providing a sensible description of the rules; addressing nonintuitiveness requires providing a satisfying explanation for why the rules are what they are. Existing laws like the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA),
and the General Data Protection Regulation (GDPR), as well as techniques within machine learning, are focused almost entirely on the problem of inscrutability. While such techniques could allow a machine learning system to comply with existing law, doing so may not help if the goal is to assess whether the basis for decision-making is normatively defensible.

In most cases, intuition serves as the unacknowledged bridge between a descriptive account and a normative evaluation. But because machine learning is often valued for its ability to uncover statistical relationships that defy intuition, relying on intuition is not a satisfying approach. This Article thus argues for other mechanisms for normative evaluation. To know why the rules are what they are, one must seek explanations of the process behind a model’s development, not just explanations of the model itself.

INTRODUCTION

I. INSCRUTABLE AND NONINTUITIVE
   A. Secret .................................................................
   B. Requiring Specialized Knowledge ...................
   C. Inscrutable ......................................................
   D. Nonintuitive ...................................................

II. LEGAL AND TECHNICAL APPROACHES TO INSCRUTABILITY ....
    A. Legal Requirements for Explanation ..............
       1. FCRA, ECOA, and Regulation B ..............
       2. GDPR ....................................................
    B. Interpretability in Machine Learning ............
       1. Purposefully Building Interpretable Models ..
       2. Post Hoc Methods ...................................
       3. Interactive Approaches ............................

III. FROM EXPLANATION TO INTUITION ..........................
    A. The Value of Opening the Black Box ............
       1. Explanation as Inherent Good ...............
       2. Explanation as Enabling Action .............
       3. Explanation as Exposing a Basis for Evaluation ..
    B. Evaluating Intuition ...................................

IV. DOCUMENTATION AS EXPLANATION ........................
    A. The Information Needed to Evaluate Models ....
    B. Providing the Necessary Information ..........

CONCLUSION ..............................................................
There can be no total understanding and no absolutely reliable test of understanding.

—Joseph Weizenbaum, “Contextual Understanding by Computers”¹

INTRODUCTION

Algorithms increasingly inform consequential decisions about our lives, with only minimal input from the people they affect and little to no explanation as to how they work.² This worries people, and rightly so. The results of these algorithms can be unnerving,³ unfair,⁴ unsafe,⁵ unpredictable,⁶ and unaccountable.⁷ How can decision makers who use algorithms be held to account for their results?

It is perhaps unsurprising that, faced with a world increasingly dominated by automated decision-making, advocates, policymakers, and legal scholars would call for machines that can explain themselves.⁸ People have a natural

---

¹. 10 COMM. ACM 474, 476 (1967). In the 1960s, the project of artificial intelligence (AI) was largely to mimic human intelligence. Weizenbaum was therefore actually arguing that computers will never fully understand humans. The purpose of AI research has changed drastically today, but there is a nice symmetry in the point that humans will never have total understanding of computers.


⁸. See infra Part II.
feel for explanation. We know how to offer explanations and can often agree when one is good, bad, in-between, on point, or off topic. Lawyers use explanation as their primary tradecraft: judges write opinions, administrators respond to comments, litigators write briefs, and everyone writes memos. Explanations are the difference between a system that vests authority in lawful process and one that vests it in an unaccountable person.9

Although we comfortably use explanations, asking someone to define the concept will often generate a blank look in response. Analytically, explanation is infinitely variable, and there can be many valid explanations for a given phenomenon or decision. Thus far, in both law and machine learning, the scholarly discourse around explanation has primarily revolved around two questions: Which kinds of explanations are most useful, and which are technically available?10 Yet, these are the wrong questions or, at least, the wrong stopping points.

Explanations of technical systems are necessary but not sufficient to achieve law and policy goals, most of which are concerned not with explanation for its own sake, but with ensuring that there is a way to evaluate the basis of decision-making against broader normative constraints such as antidiscrimination or due process. It is therefore important to ask how exactly people engage with those machine explanations in order to connect them to the normative questions of interest to law.

This Article argues that scholars and advocates who seek to use explanation to enable justification of machine learning models are relying on intuition to connect the explanation to normative concerns. Intuition is both powerful and dangerous. While this mode of justifying decision-making remains important, we must understand the benefits and weaknesses of connecting machine explanation to intuitions. Remedyng the limitations of intuition requires considering alternatives, which include institutional processes, documentation, and access to those documents.

This Article proceeds in four parts. Part I examines the various anxieties surrounding the use of automated decision-making. After discussing secrecy, lack of transparency, and lack of technical expertise, Part I argues that the two distinct, but similar, concepts that truly set machine learning decision-making apart are inscrutability and nonintuitiveness.

Part II examines laws and machine learning techniques designed specifically to address the problem of inscrutable decisions. On the legal side, Part II.A discusses the “adverse action notices” required by federal credit laws11 and the informational requirements of the European Union’s General Data Protection Regulation (GDPR).12 On the technical side, Part

10. See infra Part III.
II.B discusses various techniques used by computer scientists to make machine learning models interpretable, including designing for simplicity, approximating complex models in simpler form, extracting the most important factors in a particular decision, and allowing some degree of interaction with the models to see how changes in inputs affect outputs. These techniques can be useful in meeting the requirements of the law, but such explanations, even when they comply with the law, may be of limited practical utility.

Part III builds the connection between explanation and intuition before evaluating the merits of an intuition-centered approach to justification. It canvasses reasons besides justification that one might want explainable machines—dignity or autonomy on the one hand and consumer or data-subject education on the other—before concluding that neither is adequate to fully address the concerns with automated decision-making. Interrogating the assumptions behind a third reason—that explanation will reveal problems with the basis for decision-making—demonstrates the reliance on intuition. The remainder of Part III examines the value and limitations of intuition. With respect to machine learning in particular, although intuition can root out obviously good or bad cases, it cannot capture the cases that give machine learning its greatest value: true patterns that exceed human imagination. These cases are not obviously right or wrong, but simply strange.

Part IV aims to provide another way. Once outside the black box, all that is left is to question the process surrounding its development and use. There are large parts of the process of machine learning that do not show up in a model but can contextualize its operation, such as paths considered but not taken and the constraints that influence these choices. Where intuition is insufficient to determine whether the model’s rules are reasonable or rest on valid relationships, justification can sometimes be achieved by demonstrating and documenting due care and thoughtfulness.

I. INSCRUTABLE AND NONINTUITIVE

Scholarly and policy debates about regulating a world controlled by algorithms have been mired in difficult questions about how to observe, access, audit, or understand those algorithms.13 The difficulty has been attributed to a diverse set of problems, specifically that algorithms are

---

“secret”14 and “opaque”15 “black boxes”16 that are rarely, if ever, made “transparent”;17 that they operate on the basis of correlation rather than “causality”18 and produce “predictions”19 rather than “explanations”;20 that their behavior may lack “intelligibility”21 and “foreseeability”;22 and that they challenge established ways of being “informed”23 or “knowing.”24 These terms are frequently used interchangeably or assumed to have overlapping meanings. For example, opacity is often seen as a synonym for secrecy,25 an antonym for transparency,26 and, by implication, an impediment to understanding.27 Yet the perceived equivalence of these terms has obscured important differences between distinct problems that frustrate attempts at regulating algorithms—problems that must be disentangled before the question of regulation can even be addressed.

This Part argues that many of these challenges are not unique to algorithms or machine learning. We seek here to parse the problems raised by machine learning models more precisely and argue that they have little to do with the fact that their very existence may be unknown, that their inner workings may be opaque, or that an understanding of their operations may require specialized knowledge. What sets machine learning models apart from other decision-making mechanisms are their inscrutability and nonintuitiveness.

17. See, e.g., Citron & Pasquale, supra note 7, at 27; Tal Z. Zarsky, Transparent Predictions, 2013 U. ILL. L. REV. 1503, 1506.
18. See, e.g., Kim, supra note 4, at 875.
21. See, e.g., Brennan-Marquez, supra note 19, at 1253.
22. See, e.g., Karnow, supra note 6, at 52.
We adapt and extend a taxonomy first proposed by Jenna Burrell,28 where our primary purpose is to emphasize these last two properties and clear up confusion.29 Inscrutability and nonintuitiveness have been conflated in the past: where the property of inscrutability suggests that models available for direct inspection may defy understanding, nonintuitiveness suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition.30

A. Secret

The first common critique of algorithmic decision-making is secrecy. Secrecy captures two related, but distinct, concerns: (1) secrecy of the model’s existence and (2) secrecy of its operation.

The first concern is as old as the original Code of Fair Information Practices (FIPs), the conceptual basis for the majority of privacy laws:31 “There must be no personal-data record-keeping systems whose very existence is secret.”32 This principle underlies more recent calls to “end secret profiling” involving algorithms and machine learning, where secrecy is understood as a purposeful attempt to maintain ignorance of the very fact of profiling.33

While such worries are particularly pronounced when the government engages in algorithmic decision-making,34 similar objections arise in the commercial sector, where there are a remarkable number of scoring systems

---

28. See generally id.
30. We intentionally use the term “nonintuitive” rather the word “unintuitive” or “counterintuitive.” In our view, “unintuitive” implies a result that would not be expected but is easily understood once explained, and “counterintuitive” suggests a phenomenon that is opposite one’s expectations. Instead, we intend to refer to a phenomenon about which intuitive reasoning is not possible.
of which consumers are simply unaware.\textsuperscript{35} In many cases, this ignorance
exists because the companies engaged in such scoring are serving other businesses rather than consumers.\textsuperscript{36} But the fact that more recent forms of
hidden decision-making involve algorithms or machine learning does not change the fundamental secrecy objection—that affected parties are not
aware of the existence of the decision-making process.\textsuperscript{37}

The second secrecy concern arises where the existence of a decision-
making process is known, but its actual operation is not. Affected parties
might be aware that they are subject to such decision-making but have limited or no knowledge of how the decision-making process works.\textsuperscript{38} Among the
many terms used to describe this situation, “opacity” seems most apt, as there is enough visibility to see that the model exists but not enough to discern any of its details.

While this is perhaps the most frequent critique of algorithms and machine learning—that their inner workings remain undisclosed or inaccessible\textsuperscript{39}—it, too, has little to do with the technology specifically. It is an objection to being subject to a decision where the basis of decision-making remains secret, which is a situation that can easily occur without algorithms or machine learning.\textsuperscript{40}

There are sometimes valid reasons for companies to withhold details about a decision-making process. Where a decision-making process holds financial and competitive value and where its discovery entails significant investment or ingenuity, firms may claim protection for its discovery as a trade secret.\textsuperscript{41} Trade-secret protection applies only when firms purposefully restrict disclosure of proprietary methods,\textsuperscript{42} which creates incentives for firms to maintain secrecy around the basis for decision-making. If the use of algorithms or machine learning uniquely increases up-front investment or competitive advantage, then the incentives to restrict access to the details of


\textsuperscript{37} See Sec’y’s Advisory Comm. on Automated Pers. Data Sys., supra note 32, at 29 (discussing the lack of awareness of record keeping and use of personal data).

\textsuperscript{38} This could refer to secrecy around what data is considered or how it is used. See infra Part II.A for a discussion of these concerns with respect to the Fair Credit Reporting Act.

\textsuperscript{39} See, e.g., Robert Brauneis & Ellen P. Goodman, Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tech. 103, 107–08 (2018); Citron & Pasquale, supra note 7, at 10–11. See generally Pasquale, supra note 16.

\textsuperscript{40} See, e.g., Daniel J. Solove, Privacy and Power: Computer Databases and Metaphors for Information Privacy, 53 Stan. L. Rev. 1393, 1407, 1410 (2001) (discussing the private database industry and corporate decision-making based on consumer data).


\textsuperscript{42} Pasquale, supra note 14, at 237.
the decision-making process might be understood as peculiar to algorithms or machine learning. But if other attempts to develop decision-making processes without algorithms or machine learning involve similar costs and competitive advantage, then there is nothing special about the relationship between these technologies, trade secrets, and resistance to disclosure.43

Firms may also reject requests for further details about the basis for decision-making if they anticipate that such details may enable strategic manipulation, or “gaming,” of the inputs to the decision-making process.44 If the costs of manipulating one’s characteristics or behavior are lower than the expected benefits, rational actors would have good incentive to do so.45 Yet these dynamics, too, apply outside algorithms and machine learning; in the face of some fixed decision procedure, people will find ways to engage in strategic manipulation. The question is whether decision procedures developed with machine learning are easier or harder to game than those developed using other methods—this is not a question that can be answered in general.

B. Requiring Specialized Knowledge

One common approach to ensuring accountability for software-reliant decision-making is to require the disclosure of the underlying source code.46 While such disclosure might seem helpful in figuring out how automated decisions are rendered, the ability to make sense of the disclosed source code will depend on one’s level of technical literacy. Some minimal degree of training in computer programming is necessary to read code, although even that might not be enough.47 The problem, then, is greater than disclosure; in


45. Whether such manipulation is even possible will vary from case to case, depending on the degree to which the decision considers immutable characteristics and nonvolitional behavior. At the same time, it is unclear how easily one could even change the appearance of one’s characteristics without genuinely changing those characteristics in the process. Altering behavior to game the system might involve adjustments that actually change a person’s likelihood of having the sought-after quality or experiencing the event that such behavior is meant to predict. To the extent that “gaming” is a term used to describe validating rather than defeating the objectives of a decision system, this outcome should probably not be considered “gaming” at all. See Bambauer & Zarsky, supra note 44.


47. Desai & Kroll, supra note 46, at 5 (“[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code.”); Kroll et al., supra note 7, at 647.
the absence of the specialized knowledge required to understand source code, disclosure may offer little value to affected parties and regulators.

As Mike Ananny and Kate Crawford have observed, “Transparency concerns are commonly driven by a certain chain of logic: observation produces insights which create the knowledge required to govern and hold systems accountable.” The process of moving from observation to knowledge to accountability cannot be assumed; in many cases, the ability to leverage observations for accountability requires preexisting knowledge that allows observers to appreciate the significance of a disclosure. Transparency of systems of decision-making is important, but incomplete. But while cultivating the knowledge necessary to read source code requires time and effort, the problem of expertise—like the problem of secrecy—is not unique to algorithms.

C. Inscrutable

Rather than programming computers by hand with explicit rules, machine learning relies on pattern-recognition algorithms and a large set of examples to uncover relationships in the data that might serve as a reliable basis for decision-making. The power of machine learning lies not only in its ability to relieve programmers of the difficult task of producing explicit instructions for computers, but in its capacity to learn subtle relationships in data that humans might overlook or cannot recognize. This power can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse.

We define this difficulty as “inscrutability”—a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension. While there is a long history to such concerns, evidenced most obviously by the term “byzantine,” the complexity of rules that result from machine learning can far exceed those of the most elaborate bureaucracy. The challenge in such circumstances is not a lack of awareness, disclosure, or expertise, but the sheer scope and sophistication of the model.

Intuitively, complexity would seem to depend on the number of rules encoded by a model, the length of a rule (i.e., the number of factors that figure into the rule), and the logical operations involved in the rule. These properties, however, can be specified more precisely. Four mathematical

---

properties related to model complexity are linearity, monotonicity, continuity, and dimensionality.

A linear model is one in which there is a steady change in the value of the output as the value of the input changes. Linear models tend to be easier for humans to understand and interpret because the relationship between variables is stable and lends itself to straightforward extrapolation. In contrast, the behavior of nonlinear models can be far more difficult to predict, even when they involve simple mathematical operations like exponential growth.

A monotonic relationship between variables is a relationship that is either always positive or always negative. That is, for every change in input value, the direction of the corresponding change in output value will remain consistent, whether an increase or decrease. Monotonicity aids interpretability because it too permits extrapolation and guarantees that the value of the output only moves in one direction. If, however, the value of the output goes up and down haphazardly as the value of the input moves steadily upward, the relationship between variables can be difficult to grasp or predict.

Discontinuous models include relationships where changes in the value of one variable do not lead to a smooth change in the associated value of another. Discontinuities can render models far less intuitive because they make it impossible to think in terms of incremental change. A small change in input may typically lead to small changes in outputs, except for occasional and seemingly arbitrary large jumps.

The dimensionality of a model is the number of variables it considers. Two-dimensional models are easy to understand because they can be visualized graphically with a standard plot (with the familiar \( x \) and \( y \) axes). Three-dimensional models also lend themselves to effective visualization (by adding a \( z \) axis), but humans have no way to visualize models with more than three dimensions. While people can grasp relationships between multiple

---

54. Mathematically, this means that the function is described by a constant slope, which can be represented by a line. Yin Lou et al., *Intelligible Models for Classification and Regression*, in *PROCEEDINGS OF THE 18TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING* 150, 150 (2012).

55. See id. at 151.


58. See id.

59. See id.

60. See *Continuous Function*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014) (noting that a continuous function does not suddenly jump at a given point or take widely differing values arbitrarily close to that point).


62. See *Dimension (Dimensionality)*, A DICTIONARY OF COMPUTER SCIENCE (7th ed. 2016).

63. See *Cartesian Plane*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

64. See *Cartesian Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014); *n-Dimensional Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).
variables without the aid of a graph, we will struggle to understand the full set of relationships that the model has uncovered as the number of dimensions grows. The more variables that the model includes, the more difficult it will be to keep all the interactions between variables in mind and thus predict how the model would behave given any particular input.65

In describing how these properties of models might frustrate human understanding, we have relied on terms like intuition, extrapolation, and prediction. The same cognitive capacity underlies all three: mentally simulating how a model turns inputs into outputs.66 As computer scientist Zachary Lipton explains, simulatability—the ability to practically execute a model in one’s mind—is an important form of understanding a model.67 Such simulations can be either complete or partial. In the former, a person is able to turn any combination of inputs into the correct outputs, while in the latter, understanding might be limited to the relationships between a subset of input and output variables (i.e., how changes in certain inputs affect the output).

Simulation is a remarkably flat and functional definition of understanding, but it seems like a minimum requirement for any more elaborate definition.68 This notion of understanding has nothing to say about why the model behaves the way it does; it is simply a way to account for the facility with which a person can play out how a model would behave under different circumstances. When models are too complex for humans to perform this task, they have reached the point of inscrutability.

D. Nonintuitive

A different line of criticism has developed that takes issue with disclosures that reveal some basis for decision-making that defies human intuition about the relevance of certain variables.69 The problem in such cases is not

65. See Lehr & Ohm, supra note 51, at 700.
67. Id.
68. While we limit our discussion to simulatability, inscrutability is really a broader concept. In particular, models might be difficult to understand if they consider features or perform operations that do not have some ready semantic meaning. Burrell, supra note 15, at 10. For example, a deep-learning algorithm can learn on its own which features in an image are characteristic of different objects (the standard example being cats). Bornstein, supra note 2. Part III.A.3, infra, returns to one such example that involves distinguishing between wolves and huskies. See infra notes 246–47 and accompanying text. An algorithm will usually learn to detect edges that differentiate an object from its background, but it might also engineer features on its own that have no equivalent in human cognition and therefore defy description. See Lipton, supra note 66, at 98 (discussing decomposability). This aspect of inscrutability, however, is of slightly less concern for this Article. Most methods that are common in the kinds of applications that apportion important opportunities (e.g., credit) involve features that have been handcrafted by experts in the domain (e.g., length of employment) and accordingly will usually not face this problem. See infra note 120 and accompanying text.
inscrutability, but an inability to weave a sensible story to account for the statistical relationships in the model. Although the statistical relationship that serves as the basis for decision-making might be readily identifiable, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision. As Paul Ohm explains:

We are embarking on the age of the impossible-to-understand reason, when marketers will know which style of shoe to advertise to us online based on the type of fruit we most often eat for breakfast, or when the police know which group in a public park is most likely to do mischief based on the way they do their hair or how far from one another they walk.

Even though it is clear which statistical relationships serve as the basis for decision-making in this case, why such statistical relationships exist is mystifying. This is a crucial and consistent point of confusion. The demand for intuitive relationships is not the demand for disclosure or accessible explanations; it is a demand that decision-making rely on reasoning that comports with intuitive understanding of the phenomenon in question. In social science, similar expectations are referred to as “face validity”—the subjective sense that some measure is credible because it squares with our existing understanding of the phenomenon. While such demands are not unique to algorithms and machine learning, the fact that such computational tools are designed to uncover relationships that defy human intuition explains why the problem will be particularly pronounced in these cases.

Critics have pinned this problem on the use of “[m]ere correlation” in machine learning, which frees it to uncover reliable, if incidental, relationships in the data that can then serve as the basis for consequential decision-making. Despite being framed as an indictment of correlational analysis, however, it is really an objection to decision-making that rests on particular correlations that defy familiar causal stories—even though these stories may be incorrect. This has led to the mistaken belief that forcing

---


70. See Brennan-Marquez, supra note 19, at 1280–97.
72. Id.
74. Kim, supra note 4, at 875, 883.
75. Id.; see also James Grimmelmamn & Daniel Westreich, Incomprehensible Discrimination, 7 CALIF. L. REV. ONLINE 164, 173 (2016).
76. See Brennan-Marquez, supra note 19, at 1280–97.
77. See DANIEL KAHNEMAN, THINKING FAST AND SLOW 199–200 (2011) (discussing the “narrative fallacy”); id. at 224 (“Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not.”).
decision-making to rest on causal mechanisms rather than mere correlations will ensure intuitive models.\textsuperscript{78}

Causal relationships can be exceedingly complex and nonintuitive, especially when dealing with human behavior.\textsuperscript{79} Indeed, causal relationships uncovered through careful experimentation can be as elaborate and unexpected as the kinds of correlations uncovered in historical data with machine learning.\textsuperscript{80} If we consider all the different factors that cause a person to take an action—mood, amount of sleep, food consumption, rational choice, and many other things—it quickly becomes clear that causality is not particularly straightforward.\textsuperscript{81} The only advantage of models that rely on causal mechanisms in such cases would be the reliability of their predictions (because the models would be deterministic rather than probabilistic), not the ability to interrogate whether the identified causal relationships comport with human intuitions and values. Given that much of the interest in causality stems from an unwillingness to simply defer to predictive accuracy as a justification for models, improved reliability will not be a satisfying answer.

* * *

The demand for intuitive relationships reflects a desire to ensure that there is a way to assess whether the basis of decision-making is sound, as a matter of validity and as a normative matter. We want to be able to do more than simply simulate a model; we want to be able to \textit{evaluate} it. One way to ensure this possibility is to force a model to rely exclusively on features that bear a manifest relationship to the outcome of interest, on the belief that well-justified decisions are those that rest on relationships that conform to familiar and permissible patterns.

Achieving this type of intuitiveness requires addressing inscrutability as a starting point. An understandable model is necessary because there can be nothing intuitive about a model that resists all interrogation. But addressing inscrutability is not sufficient. A simple, straightforward model might still defy intuition if it has not been constrained to only use variables with an intuitive relationship to the outcome.\textsuperscript{82}

\textsuperscript{78} These critiques also presume that causal mechanisms that exhaustively account for the outcomes of interest actually exist (e.g., performance on the job, default, etc.), yet certain phenomena might not be so deterministic; extrinsic random factors may account for some of the difference in the outcomes of interest. Jake M. Hofman, Amit Sharma & Duncan J. Watts, \textit{Prediction and Explanation in Social Systems}, 355 \textit{Science} 486, 488 (2017).

\textsuperscript{79} Id.

\textsuperscript{80} See id.

\textsuperscript{81} Attempts to model causation require limiting the features considered as potential causes because, to a certain extent, almost any preceding event could conceivably be causally related to the later one. Judea Pearl, \textit{Causality: Models, Reasoning and Inference} 401–28 (2d ed. 2009).

\textsuperscript{82} See, e.g., Jiaming Zeng, Berk Ustun & Cynthia Rudin, \textit{Interpretable Classification Models for Recidivism Prediction}, 180 J. \textit{Royal Stat. Soc’y} 689 (2017). Note that in this and related work, the researchers limit themselves to features that are individually and intuitively related to the outcome of interest. See id. at 693–97. If these methods begin with features that do not have such a relationship, the model might be simple enough to inspect but too strange to square with intuition. See infra Part III.B.
Insisting on intuitive relationships is not the only way to make a model evaluable. To the extent that intuitiveness is taken to be an end in itself rather than a particular means to the end of ensuring sound decision-making, its proponents risk overlooking other, potentially more effective, ways to achieve the same goal. The remainder of this Article considers the different paths we might take to use explanations of machine learning models to regulate them.

II. LEGAL AND TECHNICAL APPROACHES TO INSCRUTABILITY

This moment is not the first time that law and computer science have attempted to address algorithmic decision-making with explanation requirements. Credit scoring has long been regulated, in part, by requiring “adverse action notices,” which explain adverse decisions to consumers. In Europe, concern about automated decisions has been a neglected part of data protection law for more than two decades, with interest in them reinvigorated by the GDPR. On the machine learning side, the subfield of “interpretability”—within which researchers have been attempting to find ways to understand complex models—is over thirty years old.

What seems to emerge from the law and computer science is a focus on two kinds of explanation. The first concerns accounting for outcomes—how particular inputs lead to a particular output. The second concerns the logic of decision-making—full or partial descriptions of the rules of the system. This Part reviews the legal and technical approaches to outcome and logic-based explanations.

A. Legal Requirements for Explanation

Though much of the current concern over inscrutable systems stems from the growing importance of machine learning, inscrutable systems predate this technique. As a result, regulations that require certain systems to explain themselves already exist. This section discusses two examples of legal systems and strategies that rely on different types of explanations: credit reporting statutes, which rely on outcome-based explanations, and the GDPR, which mandates logic-based explanations. Credit scoring predates machine learning, and is governed by two statutes: the Fair Credit Reporting Act (FCRA) and the Equal Credit Opportunity Act (ECOA). Statistical credit-scoring systems take information about consumers as inputs, give the

83. See infra notes 100–01 and accompanying text.
inputs certain point values, add them to obtain a total score, and then make
decisions based on that score. Each of these statutes require “adverse action
notices” that must include a statement of reasons for denials of credit or other
credit-based outcomes.\textsuperscript{88} This is an example of what we call outcome-based
explanations: a description of the facts that proved relevant to a decision, but
not a description of the decision-making rules themselves.

Articles 13–15 of the GDPR require data subjects to have access to
“meaningful information about the logic involved” in any automated
decision-making that significantly affects them.\textsuperscript{89} As the law is still new, the
import and proper interpretation of this requirement remain unclear. In
advance of a definitive interpretation, the GDPR appears to ask for a
functional description of the model—enough of a description of the rules
governing decision-making such that a data subject can vindicate her
substantive rights under the GDPR and human rights laws.\textsuperscript{90} This is an
example of logic-based explanations: a description of the reasoning behind
a decision, not just the relevant inputs to the decision.

1. FCRA, ECOA, and Regulation B

The most straightforward legal requirement to explain inscrutable
decision-making is the adverse action notice. In 1970, Congress passed
FCRA\textsuperscript{91} to begin to rein in the unregulated credit industry. FCRA was “the
first information privacy legislation in the United States.”\textsuperscript{92} It limits to whom
and for what purposes credit reports can be disclosed,\textsuperscript{93} allows consumers
access to their credit reports,\textsuperscript{94} and requires credit reporting agencies
(CRAs)—for example, Experian, Transunion, and Equifax—to employ
procedures to ensure accuracy and govern dispute resolution.\textsuperscript{95} FCRA was
not initially concerned with how decisions were made, but rather with the
then-new phenomenon of amassing large quantities of information.\textsuperscript{96} Four
years later, however, Congress passed ECOA\textsuperscript{97} and took aim at the decision-

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{88} 15 U.S.C. §§ 1681m, 1691d(2).
\item \textsuperscript{89} GDPR, supra note 12, arts. 13(f)(2), 14(g)(2), 15(1)(h) (requiring access to
“meaningful information about the logic” of automated decisions).
\item \textsuperscript{90} See Andrew D. Selbst & Julia Powles, Meaningful Information and the Right to
Explaination, 7 INT’L DATA PRIVACY L. 233, 256 (2017). There is a vigorous debate in the
literature about the “right to explanation” in the GDPR. See infra notes 143–45 and
accompanying text. As a discussion of positive law, this debate is connected to, but different
than, the point we seek to make about the GDPR—that it is one example of a law that operates
by asking for the logic of a system. Even if there is held to be no “right to explanation” in the
GDPR, one could imagine an equivalent law that encodes such a requirement.
\item \textsuperscript{91} Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1127 (1970) (codified as
\item \textsuperscript{92} Priscilla M. Regan, Legislating Privacy: Technology, Social Values, and
\item \textsuperscript{93} 15 U.S.C. § 1681b.
\item \textsuperscript{94} Id. § 1681g.
\item \textsuperscript{95} Id. §§ 1681e(b), 1681i.
\item \textsuperscript{96} 115 Cong. Rec. 2410 (1969).
\item \textsuperscript{97} Equal Credit Opportunity Act, Pub. L. No. 93-495, 88 Stat. 1521 (1974) (codified as
\end{itemize}
\end{footnotesize}
making process. ECOA prohibits discrimination in credit decisions on the basis of race, color, religion, national origin, sex, marital status, age (for adults), receipt of public assistance income, or exercise in good faith of the rights guaranteed under the Consumer Credit Protection Act.

ECOA introduced the adverse action notice requirement. When a creditor takes an adverse action against an applicant, the creditor must give a statement of “specific reasons” for the denial. When FCRA later adopted a similar requirement, it expanded the notice to cover uses of credit information beyond decisions made by creditors, including the use of such information in employment decisions.

ECOA’s notice requirement was implemented by the Federal Reserve Board via Regulation B, which mandates that the “statement of reasons . . . must be specific and indicate the principal reason(s) for the adverse action.” The regulation also notes that it is insufficient to “state[] that the adverse action was based on the creditor’s internal standards or policies or that the applicant . . . failed to achieve a qualifying score on the creditor’s credit scoring system.” An appendix to Regulation B offers a sample notification form designed to satisfy both the rule’s and FCRA’s notification requirements. Sample Form 1 offers twenty-four reason codes, including such varied explanations as “no credit file,” “length of employment,” or “income insufficient for amount of credit requested.”

---

98. Id. § 502, 88 Stat. at 1521 (noting that the purpose of the legislation is to ensure credit is extended fairly, impartially, and without regard to certain protected classes).
100. Id. § 1691(d)(2)(B); Winnie F. Taylor, Meeting the Equal Credit Opportunity Act’s Specificity Requirement: Judgmental and Statistical Scoring Systems, 29 BUFF. L. REV. 73, 82 (1980) (“For the first time, federal legislation afforded rejected credit applicants an automatic right to discover why adverse action was taken.”).
105. Id.
106. 12 C.F.R. pt. 1002, app. C (2018). The form’s listed options are:
- Credit application incomplete
- Insufficient number of credit references provided
- Unacceptable type of credit references provided
- Unable to verify credit references
- Temporary or irregular employment
- Unable to verify employment
- Length of employment
- Income insufficient for amount of credit requested
- Excessive obligations in relation to income
- Unable to verify income
- Length of residence
- Temporary residence
- Unable to verify residence
- No credit file
- Limited credit experience
- Poor credit performance with us
- Delinquent past or present credit obligations with others
- Collection action or judgment
necessary to use the form, most creditors tend to report reasons contained on that form because there is a safe harbor for “proper use” of the form.107

Adverse action notices aim to serve three purposes: (1) to alert a consumer that an adverse action has occurred;108 (2) to educate the consumer about how such a result could be changed in the future;109 and (3) to prevent discrimination.110 As the rest of this section will show, these are commonly cited reasons for relying on explanations as a means of regulation as a general matter. The first rationale, consumer awareness, is straightforward enough. It is a basic requirement of any information-regulation regime that consumers be aware of systems using their information.111 But the relationship between adverse action notices and the other two rationales—consumer education and antidiscrimination—requires further exploration.

Adverse action notices can be helpful for consumer education. As Winnie Taylor pointed out shortly after the passage of ECOA, some reasons—“no credit file” and “unable to verify income”—are self-explanatory and would allow a consumer to take appropriate actions to adjust.112 Conversely, some explanations, such as “length of employment” and home ownership, are harder to understand or act on.113 This suggests that an explanation of a specific decision may be informative, but it may not reveal an obvious path to an alternative outcome.

There are also situations in which it may not even be informative. Taylor imagined a hypothetical additive credit-scoring system with eight different features—including whether an applicant owns or rents, whether he has a home phone, and what type of occupation he has, among other things—each assigned different point values.114 In a system like that, someone who comes up one point short could find himself with every factor listed as a “principal

<table>
<thead>
<tr>
<th>Garnishment or attachment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreclosure or repossession</td>
</tr>
<tr>
<td>Bankruptcy</td>
</tr>
<tr>
<td>Number of recent inquiries on credit bureau report</td>
</tr>
<tr>
<td>Value or type of collateral not sufficient</td>
</tr>
<tr>
<td>Other, specify: _________</td>
</tr>
</tbody>
</table>


109. Id. (“[R]ejected credit applicants will now be able to learn where and how their credit status is deficient and this information should have a pervasive and valuable educational benefit. Instead of being told only that they do not meet a particular creditor’s standards, consumers particularly should benefit from knowing, for example, that the reason for the denial is their short residence in the area, or their recent change of employment, or their already over-extended financial situation.”).

110. Id. (“The requirement that creditors give reasons for adverse action is . . . a strong and necessary adjunct to the antidiscrimination purpose of the legislation, for only if creditors know they must explain their decisions will they effectively be discouraged from discriminatory practices.”).

111. See supra note 32 and accompanying text.

112. Taylor, supra note 100, at 97.

113. Id. at 95.

114. Id. at 105–07.
reason”115 for the denial. In one sense, this must be correct because a positive change in any factor at all would change the outcome. In another sense, however, choosing arbitrarily among equivalently valid reasons runs counter to the instruction to give specific and actionable notice.

Taylor also described a real system from that era, complex in all the various ways described in Part I—nonlinear, nonmonotonic, discontinuous, and multidimensional:

[A]pplicants who have lived at their present address for less than six months are awarded 39 points, a level which they could not reach again until they had maintained the same residence for seven and one-half years. Furthermore, applicants who have been residents for between six months and 1 year 5 months (30 points) are considered more creditworthy than those who have been residents for between 1 and 1/2 years and 3 years 5 months (27 points).116

If the creditor tried to explain these rules simply, it would leave information out, but if the creditor were to explain in complete detail, it would likely overwhelm a credit applicant. This is an equivalent problem to simply disclosing how a model works under the banner of transparency; access to the model is not the same as understanding.117

The Federal Reserve Board recognized this problem, observing that, although all the principal reasons must be disclosed, “disclosure of more than four reasons is not likely to be helpful to the applicant.”118 The difficulty is that there will be situations where complexity cannot be avoided in a faithful representation of the scoring system, and listing factors alone will fail to accurately explain the decision, especially when the list is limited to four.119 It is worth noting that modern credit systems appear not to be based on such complex models,120 likely due to the very existence of FCRA and ECOA. Credit predictions tend to rely on features that bear an intuitive relationship to default, such as past payment history.121 But the point is more general:

115. See supra note 104 and accompanying text.
116. Taylor, supra note 100, at 123.
117. See Ananny & Crawford, supra note 24, at 979 (“Transparency can intentionally occlude.”).
119. The document also states that the “specific reasons . . . must relate to and accurately describe the factors actually considered or scored by a creditor . . . . A creditor need not describe how or why a factor adversely affected an applicant . . . . If a creditor bases the . . . adverse action on a credit scoring system, the reasons disclosed must relate only to those factors actually scored in the system.” 12 C.F.R. pt. 1002 supp. I, para. 9(b)(2).
approaches based on giving specific reasons for outcomes can fail where the system is too complex. The adverse action notice fares worse as an antidiscrimination measure. By 1974, forcing hidden intentions into the open was a common technique for addressing discrimination.\textsuperscript{122} Just one year before ECOA’s passage, \textit{McDonnell Douglas Corp. v. Green}\textsuperscript{123} laid out the canonical Title VII burden-shifting framework for disparate treatment, which requires a defendant to rebut a prima facie case of employment discrimination with a nondiscriminatory reason and gives plaintiffs a chance to prove that the proffered reason is pretextual.\textsuperscript{124} Just two years before that, the U.S. Supreme Court in \textit{Griggs v. Duke Power Co.}\textsuperscript{125} recognized disparate impact doctrine.\textsuperscript{126} Disparate impact attributes liability for a facially neutral decision that has a disproportionate adverse effect on a protected class unless the decision maker can provide a legitimate business reason for the decision and no equally effective but less discriminatory alternative exists.\textsuperscript{127} Its initial purpose was arguably to smoke out intentional discrimination where intent was hidden.\textsuperscript{128} Thus, ECOA pursued the same goal—to prevent discrimination by forcing decision-making into the open.

While forcing stated reasons into the open captures the most egregious forms of intentional discrimination, it does not capture much else. Although, in some cases, Regulation B bars collection of protected-class information,\textsuperscript{129} race, gender, and other features can be reliably inferred from sufficiently rich datasets.\textsuperscript{130} Should creditors seek to discriminate intentionally by considering membership in a protected class, they would have to affirmatively lie about such behavior lest they reveal obvious wrongdoing. This form of intentional discrimination is thus addressed by disclosure. Should creditors rely on known proxies for membership in a protected class, however, while they would have to withhold the true relevance of these features in predicting creditworthiness, they could cite them honestly as reasons for the adverse action. The notice requirement therefore does not place meaningful constraints on creditors, nor does it create additional or

\textsuperscript{123} 411 U.S. 792 (1973).
\textsuperscript{124} Id. at 805. The Supreme Court later found that a jury may presume that if all the employer had was pretext, that itself is evidence of discrimination. St. Mary’s Honor Ctr. v. Hicks, 509 U.S. 502, 511 (1993) (“The factfinder’s disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.”).
\textsuperscript{125} 401 U.S. 424 (1971).
\textsuperscript{126} Id. at 431.
\textsuperscript{127} 42 U.S.C. § 2000e-2(k)(1)(A) (2012). This description ignores the word “refuse” in the statute, but is probably the more common reading. Barocas & Selbst, \textit{supra} note 4, at 709.
\textsuperscript{129} 12 C.F.R. § 1002.5 (2018).
\textsuperscript{130} Barocas & Selbst, \textit{supra} note 4, at 692.
unique liability beyond that present in the antidiscrimination provisions of
the rest of the regulation.  

More importantly, creditors using quantitative methods that do not
expressly consider protected-class membership are likely not engaged in
intentional discrimination, yet the scoring systems might very well evince a
disparate impact. While ECOA does not expressly provide for a disparate
impact theory of discrimination, case law suggests that it is very likely
available.

The adverse action notice approach has two specific shortcomings for a
disparate impact case. First, when reviewing such a notice, the consumer
only has access to her own specific outcome. Her single point of reference
does not provide any understanding of the frequency of denials along
protected-class lines, so she cannot observe disparate impact. Absent
understanding of the logic of the system—for example, how different inputs
are weighted—she cannot even look at the decision-making to try to guess
whether it is discriminatory; the notice simply provides no basis to bring a
suit.

Second, disparate impact has a different relationship to reasons behind
decisions than does intentional discrimination. While for intentional
discrimination, a consumer only needs to know that the decision was not
made for an improper reason, knowing the specific reasons for which it was
made becomes important for a disparate impact case. That is to say, it is
not only important to understand how a statistical system converts inputs to
specific outputs, but also why the system was set up that way.

As discussed in Part I, one avenue to ensure the existence of an explanation
of why the rules are the way they are is to require that the rules be based on
intuitive relationships between input and output variables. This is the
approach advocated by several scholars, particularly those focused on
discrimination. As is discussed in Part IV, it is not the only way, but this
inability to engage with the normative purposes of the statute is a clear
shortcoming of explanations based solely on the outcome of a single case,
which provides neither the logic of the system nor any information about its
normative elements.

132. The Supreme Court has not ruled that it is available, but most circuit courts that have
considered it have permitted it. See Mikella Hurley & Julius Adebayo, Credit Scoring in the
Era of Big Data, 18 Yale J. L. & Tech. 148, 193 (2016) (citing Golden v. City of Columbus,
404 F.3d 950, 963 (6th Cir. 2005)). In addition, the Supreme Court ruled in 2015 that disparate
impact theory was cognizable in the Fair Housing Act, which also does not expressly provide
for it. Texas Dep’t of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507,
133. Barocas & Selbst, supra note 4, at 702.
134. See infra Part III.A.3.
2. GDPR

In 2016, the European Union (EU) passed the GDPR, which took effect on May 25, 2018, and replaced the 1995 Data Protection Directive. Both laws regulate automated decision-making, but in the twenty-three years of the Directive’s existence, little jurisprudence developed around that particular aspect of the law. The GDPR has created renewed interest in these provisions.

The GDPR’s discussion of automated decisions is contained in Articles 22, 13(2)(f), 14(2)(g), and 15(1)(h). Article 22 is the primary provision and states, in relevant part, the following:

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) ... (c) is based on the data subject’s explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

Articles 13–15 spell out a data subject’s right to be informed about the information that data controllers have about her. Articles 13 and 14 describe the obligations of data controllers to affirmatively notify data subjects about the uses of their information, and Article 15 delineates the access rights that data subjects have to information about how their own data is used. All three demand that the following information be available to data subjects: “the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”

135. GDPR, supra note 12, art. 99.
136. Id. art. 22(1) (“The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.”); Data Protection Directive, supra note 84, art. 15.
137. Isak Mendoza & Lee A. Bygrave, The Right Not to Be Subject to Automated Decisions Based on Profiling, in EU INTERNET LAW 77, 80–81 (2017).
138. GDPR, supra note 12, art. 22. Article 22(4) is omitted because it is not relevant to this discussion.
139. Wachter et al., supra note 23, at 89.
140. See GDPR, supra note 12, arts. 13–14.
141. See id. art. 15.
142. Id. arts. 13(2)(f), 14(2)(g), 15(1)(h).
Since passage of the GDPR, scholars have debated whether these requirements amount to a “right to explanation.” As one of us has argued elsewhere, that debate has been bogged down in proxy battles over what the phrase “right to explanation” means, but no matter whether one calls it a right to explanation, requiring that data subjects have meaningful information about the logic must mean something related to explanation. Importantly for this discussion, the Regulation demands that the “meaningful information” must be about the logic of the decisions. As we defined it in Part I, a model is inscrutable when it defies practical inspection and resists comprehension. An explanation of the logic therefore appears to precisely target inscrutability. The most important aspect of this type of explanation is that it is concerned with the operation of the model in general, rather than as it pertains to a particular outcome.

The particular type of explanation required by the GDPR will depend on the legal standards developed in the EU by the authorities charged with interpreting that law. The overall purposes of the GDPR are much broader than FCRA and ECOA. The EU treats data protection as a fundamental right, and the GDPR seeks to vindicate the following principles with respect to personal data: lawfulness, fairness, and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality; and accountability. Several of these principles are a restatement of the FIPs that have shaped privacy policy for decades.
Considered as a whole, they begin to sound like the general idea of due process in all its expansiveness.

Satisfying this requirement may in some cases involve disclosing the full set of rules behind all decision-making—that is, the entire model. But in some cases, it will not involve such radical disclosure. Depending on the specific goals at issue, the types of rules disclosed can be narrower, or the explanation can perhaps be met interactively by providing data subjects with the tools to examine how changes in their information relate to changes in outcome. One of us has argued that the GDPR’s meaningful information requirement applies “to the data subject herself” and “should be interpreted functionally and flexibly,” and that the legal standard should be that the explanation “at a minimum, enable[s] a data subject to exercise his or her rights under the GDPR and human rights law.”

Although the GDPR’s goals are broader than those of ECOA and FCRA, evaluating the ability of logic-based explanations to vindicate the goals of those statutes can demonstrate how explanations of the logic of decision-making can improve upon the shortcomings of the outcome-based approach. The three reasons were awareness, consumer (here, data subject) education, and antidiscrimination. Like in the credit domain, awareness is straightforward and encapsulated by the requirement that a data subject be made aware of the “existence” of automated decision-making. The other two rationales operate differently when logic-based explanations are provided.

Data subject education becomes more straightforward as a legal matter, if not a technical one. Absent inscrutability, a data subject would be told the rules of the model and would be able to comprehend his situation and how to achieve any particular outcome. This solves both problems that Taylor identified. Consider the system where, after the creditor totaled the point values from eight factors, a person missed on her credit application by one point. While it might be impossible to point to four factors that were “principal reasons,” the explanation of the logic—what the eight factors were, that they were all assigned point values, and that the hypothetical applicant just missed by a point—would be much more useful to that


149. The guidelines issued by the Article 29 Working Party, a body tasked with giving official interpretations of EU law, states that the full model is not required. See Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679, at 25, WP 251 (Feb. 6, 2018) (“The GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm.”). As a matter of positive law, then, this is likely to be the outcome, but in some cases it may fall short of something actually meaningful to the data subject.

150. See Selbst & Powles, supra note 90, at 236.

151. Id. at 233.

152. See supra notes 108–10 and accompanying text.

153. See supra notes 112–16 and accompanying text.
particular rejected applicant. In Taylor’s real nonlinear, nonmonotonic, discontinuous, and multidimensional example, the full complexity can be appreciated in the paragraph-long description, where a reason code would in many cases be totally unhelpful. Once machine learning enters the picture, and models become more complex, the limits on technical ability to solve inscrutability may prevent these explanations from coming to fruition. But at least in theory, explanations of the logic are sufficient for data subject education.

Turning to discrimination—which serves as a stand-in for broader normative questions about model justification—while logic-based explanations do fare better than outcome-based ones, they do not completely address the shortcomings. Any rule that is manifestly objectionable becomes visible under logic-based explanations, making them an improvement over outcome-only explanations, which shed no light on rules. This disclosure might enable one to speculate if facially neutral rules will nevertheless have a disparate impact, based on the different rates at which certain input features are held across the population. But this is ultimately little more than guesswork. Although there might not be anything about a rule that appears likely to generate a disparate impact, it still could. Alternatively, a set of rules could appear objectionable or discriminatory, but ultimately be justified. It will often be impossible to tell without more information, and the possibility of happening on a set of rules that lend themselves to intuitive normative assessment is only a matter of chance.

B. Interpretability in Machine Learning

The overriding question that has prompted fierce debates about explanation and machine learning has been whether machine learning can be made to comply with the law. As discussed in Part I, machine learning poses unique challenges for explanation and understanding—and thus challenges for meeting the apparent requirements of the law. Part II.A further demonstrated that even meeting the requirements of the law does not automatically provide the types of explanations that would be necessary to assess whether decisions are well justified. Nevertheless, addressing the potential inscrutability of machine learning models remains a fundamental step in meeting this goal.

As it happens, machine learning has a well-developed toolkit to deal with calls for explanation. There is an extensive literature on “interpretability.” Early research recognized and grappled with the challenge of explaining the decisions of machine learning models such that people using these systems

154. The Article 29 Working Party has, however, suggested that this approach is central to the “meaningful information” requirement. See Article 29 Data Protection Working Party, supra note 149, at 25.
155. See infra Part III.A.3.
156. See generally, e.g., Riccardo Guidotti et al., A Survey of Methods for Explaining Black Box Models, 51 ACM COMPUTING SURVEYS, Aug. 2018, at 1; Lipton, supra note 66.
would feel comfortable acting upon them. Practitioners and researchers have developed a wide variety of strategies and techniques to ensure that they can produce interpretable models from data—many of which may be useful for complying with existing law, such as FCRA, ECOA, and the GDPR.

Interpretability has received considerable attention in research and practice due to the widely held belief that there is a tension between how well a model will perform and how well humans will be able to interpret it. This view reflects the reasonable idea that models that consider a larger number of variables, a larger number of relationships between these variables, and a more diverse set of potential relationships is likely to be both more accurate and more complex. This will certainly be the case when the phenomenon that machine learning seeks to model is itself complex. This intuition suggests that practitioners may face a difficult choice: favor simplicity for the sake of interpretability or accept complexity to maximize performance.

While such views seem to be widely held, over the past decade, methods have emerged that attempt to sidestep these difficult choices altogether, promising to increase interpretability while retaining performance. Researchers have developed at least three different ways to respond to the demand for explanations: (1) purposefully orchestrating the machine learning process such that the resulting model is interpretable; (2) applying special techniques after model creation to approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions; and (3) providing tools that allow people to interact with the model and get a sense of its operation.

1. Purposefully Building Interpretable Models

Practitioners have a number of different levers at their disposal to purposefully design simpler models. First, they may choose to consider only a limited set of all possible variables. By limiting the analysis to a smaller set of variables, the total number of relationships uncovered in the learning process might be sufficiently limited to be intelligible to a human. It is

---

157. van Melle et al., supra note 85, at 302.
158. See, e.g., Leo Breiman, Statistical Modeling: The Two Cultures, 16 STAT. SCI. 199, 206 (2001); Lou et al., supra note 54, at 150.
159. See Breiman, supra note 158, at 208.
160. See generally id.
162. For a recent survey, see Michael Gleicher, A Framework for Considering Comprehensibility in Modeling, 4 BIG DATA 75 (2016).
163. See, e.g., id. at 81–82.
164. See, e.g., id. at 82–83.
165. See, e.g., id. at 83.
166. See id. at 81.
167. Zeng et al., supra note 82, at 690–91.
very likely that a model with five features, for example, will be more interpretable than a model with five hundred.

Second, practitioners might elect to use a learning method that outputs a model that can be more easily parsed than the output of other learning methods.\textsuperscript{168} For example, decision tree algorithms are perceived as likely to produce interpretable models because they learn nested rules that can be represented visually as a tree with subdividing branches. To understand how the model would process any particular case, practitioners need only walk through the relevant branches of the tree; to understand the model overall, practitioners can explore all the branches to develop a sense of how the model would determine all possible cases.

The experience of applying machine learning to real-world problems has led to common beliefs among practitioners about the relative interpretability of models that result from different learning methods and how well they perform. Conventional wisdom suggests that there is a trade-off between interpretability and accuracy.\textsuperscript{169} Methods like linear regression\textsuperscript{170} generate models perceived as highly interpretable, but relatively low performing, while methods like deep learning\textsuperscript{171} result in high-performing models that are exceedingly difficult to interpret.\textsuperscript{172} While researchers have pointed out that such comparisons do not rest on a rigorous definition of interpretability or empirical studies,\textsuperscript{173} such beliefs routinely guide practitioners’ decisions when applying machine learning to different kinds of problems.\textsuperscript{174}

Another method is to set the parameters of the learning process to ensure that the resulting model is not so complex that it defies human comprehension. For example, even decision trees will become unwieldy for humans if they involve an excessively large number of branches and leaves.\textsuperscript{175} Practitioners routinely set an upper bound on the number of leaves to constrain the complexity of the model.\textsuperscript{176} For decades, practitioners in regulated industries like credit and insurance have purposefully limited themselves to a relatively small set of features and less sophisticated learning methods.\textsuperscript{177} In so doing, they have been able to generate models that lend themselves to sensible explanation, but they may have forgone the increased accuracy that would result from a richer and more advanced analysis.\textsuperscript{178}

\begin{thebibliography}{99}
\bibitem{168} See Lehr & Ohm, \textit{supra} note 51, at 688–95.
\bibitem{169} See, \textit{e.g.}, Breiman, \textit{supra} note 158, at 208.
\bibitem{170} See \textit{Regression}, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).
\bibitem{172} Breiman, \textit{supra} note 158, at 206.
\bibitem{174} See Lipton, \textit{supra} note 66, at 99.
\bibitem{175} \textit{Id.} at 98.
\bibitem{176} See \textit{id.} at 99.
\bibitem{177} Hall et al., \textit{supra} note 120.
\bibitem{178} \textit{Id.}
\end{thebibliography}
Linear models remain common in industry because they allow companies to much more readily comply with the law. When they involve a sufficiently small set of features, linear models are concise enough for a human to grasp the relevant statistical relationships and to simulate different scenarios. They are simple enough that a full description of the model may amount to the kind of meaningful information about the logic of automated decisions required by the GDPR. At the same time, linear models also immediately highlight the relative importance of different features by assigning a specific numerical weight to each feature, which allows companies to quickly extract the principal factors for an adverse action notice under ECOA.

Beyond the choice of features, learning method, or learning parameters, there are techniques that can make simplicity an additional and explicit optimization criterion in the learning process. The most common such method is regularization. Much like setting an upper limit on the number of branches in a decision tree, regularization allows the learning process to factor in model complexity by assigning a cost to excess complexity. In doing so, model simplicity becomes an additional objective alongside model performance, and the learning process can be set up to find the optimal trade-off between these sometimes-competing objectives.

Finally, the learning process can also be constrained such that all features exhibit monotonicity. Monotonicity constraints are widespread in credit scoring because they make it easier to reason about how scores will change when the value of specific variables change, thereby allowing creditors to automate the process of generating the reason codes required by FCRA and ECOA. As a result of these legal requirements, creditors and other data-

---

179. Id.
180. See Lipton, supra note 66, at 98.
181. See Gleicher, supra note 162, at 81–82.
182. See id. at 81. One commonly used version of this method is Lasso. See generally Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, 58 J. ROYAL STAT. SOC’Y 267 (1996). It was originally designed to increase accuracy by avoiding overfitting, which occurs when a model assigns significance to too many features and thus accidentally learns patterns that are peculiar to the training data and not representative of real-world patterns. See id. at 267. Machine learning is only effective in practice when it successfully identifies robust patterns while also ignoring patterns that are specific to the training data. See David J. Hand, Classifier Technology and the Illusion of Progress, 21 STAT. SCI. 1, 2 (2006). Lasso increases accuracy by forcing the learning process to ignore relationships that are relatively weak, and therefore more likely to be artifacts of the training data. See Tibshirani, supra, at 268. Because Lasso works by strategically removing unnecessary features, the technique can simultaneously improve interpretability (by reducing complexity) in many real-world applications and increase performance (by avoiding overfitting). See id. at 267. As such, improved interpretability need not always decrease performance. But where potential overfitting is not a danger, regularization methods may result in degradations in performance. See Gleicher, supra note 162, at 81–82.
183. Gleicher, supra note 162, at 81.
184. Recall that monotonicity implies that an increase in an input variable can only result in either an increase or decrease in the output; it can never change from one to the other. See supra notes 57–58 and accompanying text.
185. See, e.g., Hall et al., supra note 120. Monotonicity allows creditors to rank order variables according to how much the value of each variable in an applicant’s file differs from
driven decision makers often have incentives to ensure their models are interpretable by design.

2. Post Hoc Methods

There exists an entirely different set of techniques for improved interpretability that does not place any constraints on the model-building process. Instead, these techniques begin with models learned with more complex methods and attempt to approximate them with simpler and more readily interpretable methods. Most methods in this camp generate what can be understood as a model of the model.

These methods attempt to overcome the fact that simpler learning methods cannot always reliably discover as many useful relationships in the data. For example, the learning process involved in decision trees is what is known as a “greedy algorithm.” ¹⁸⁶ Once the learning process introduces a particular branch, the method does not permit walking back up the branch. ¹⁸⁷ Therefore, relationships between items on two different branches will not be discovered. ¹⁸⁸ Despite lacking the same limitation, more complex learning methods, such as deep learning, do not result in models as interpretable as decision trees. Nonetheless, rules that cannot be learned with simpler methods can often be represented effectively by simpler models. ¹⁸⁹ Techniques like rule extraction ¹⁹⁰ allow simple models to “cheat” because the answers that simpler learning methods would otherwise miss are known ahead of time. ¹⁹¹

This approach can be costly and it does not have universal success. ¹⁹² Despite practitioners’ best efforts, replicating the performance of more complex models in a simple enough form might not be possible where the phenomena are particularly complex. For example, using a decision tree to approximate a model developed with deep learning might require too large a number of branches and leaves to be understandable in practice. ¹⁹³

When these methods work well, they ensure that the entire set of relationships learned by the model can be expressed concisely, without

¹⁸⁷. Id.
¹⁸⁸. Id. at 93 (noting that, although the greedy algorithm may find a nonoptimal solution, it will not discover relationships between unrelated branches).
¹⁸⁹. Gleicher, supra note 162, at 82.
¹⁹⁰. Rule extraction is the name for a set of techniques used to create a simplified model of a model. The technical details of their operation are beyond the scope of this paper. See generally Nahla Barakat & Andrew P. Bradley, Rule Extraction from Support Vector Machines: A Review, 74 NEUROCOMPUTING 178 (2010); David Martens et al., Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines, 183 EUR. J. OPERATIONAL RES. 1466 (2007).
¹⁹¹. Gleicher, supra note 162, at 82.
¹⁹². Id.
¹⁹³. See Lipton, supra note 66, at 98.
giving up much performance. Accordingly, they serve a similar role to the interpretability-driven design constraints discussed above. When they do not work as well, arriving at an interpretable model might necessitate sacrificing some of the performance gained by using the more complex model. But even when these methods involve a notable loss in performance, the resulting models frequently perform far better than simple methods alone.

Other tools have also emerged that attack the problem of interpretability from a different direction. Rather than attempting to ensure that machine learning generates an intelligible model overall, these new tools furnish more limited explanations that only account for the relative importance of different features in particular outcomes—similar to the reason codes required by FCRA and ECOA. At a high level, most of these methods adopt a similar approach: they attempt to establish the importance of any feature to a particular decision by iteratively varying the value of that feature while holding the value of other features constant.

These tools seem well suited for the task set by ECOA, FCRA, or other possible outcome-oriented approaches: explaining the principal reasons that account for the specific adverse decision. As we further discuss in the next section, there are several reasonable ways to explain the same specific outcome. These methods are useful for two of the most common: (1) determining the relative contribution of different features, or (2) identifying the features whose values would have to change the most to change the outcome. One could imagine applying these methods to models that consider an enormous range of features and map out an exceedingly complex set of relationships. While such methods will never make these relationships completely sensible to a human, they can provide a list of reasons that might help provide reason codes for a specific decision.
Unfortunately, these methods may not work well in cases where models take a much larger set of features into account. Should many features each contribute a small amount to a particular determination, listing each feature in an explanation is not likely to be helpful. This is the machine learning version of Taylor’s hypothetical eight-factor credit example. The number of features identified as influential might be sufficiently large that the explanation would simply reproduce the problem of inscrutability that it aims to address. The only alternative in these cases—arbitrarily listing fewer reasons than the correct number—is also unsatisfying when all features are equivalently, or nearly equivalently, important. As it happens, post hoc explanations for credit and other similarly important decisions are likely to be most attractive precisely when they do not seem to work well—that is, when the only way to achieve a certain level of performance is to vastly expand the range of features under consideration.

These methods are also unlikely to generate explanations that satisfy logic-like approaches like the GDPR. Indeed, such techniques pose a unique danger of misleading people into believing that the reasons that account for specific decisions must also apply in the same way for others—that the reasons for a specific decision illustrate a general rule. Understandably, humans tend to extrapolate from explanations of specific decisions to similar cases, but the model—especially a complex one—may have a very different basis for identifying similar-seeming cases. These methods offer explanations that apply only to the case at hand and cannot be extrapolated to decisions based on other input data.

3. Interactive Approaches

One final set of approaches is interactive rather than explanatory. Practitioners can allow people to get a feel for their models by producing interactive interfaces that resemble the methods described in the previous sections. This can take two quite different forms. One is the type proposed by Danielle Citron and Frank Pasquale and implemented, for example, by Credit Karma. Beginning with a person’s baseline credit information, Credit Karma offers a menu of potential changes, such as opening new credit cards, obtaining a new loan, or going into foreclosure. This does

200. See supra notes 114–15 and accompanying text.
202. See id.
203. See Citron & Pasquale, supra note 7, at 28–30 (discussing “interactive modeling”).
205. Id.
206. Id.
not amount to a full explanation because a person at a different starting point could make similar moves with different outcomes, but it gives the individual user a partial functional feel for the logic of the system as it applies to him specifically.

The second is more complicated and abstract. Mireille Hildebrandt has proposed something she terms “transparency-enhancing technologies.”

Such technologies would implement an interface that would allow people to simultaneously adjust the value of multiple features in a model with the goal of providing a loose sense of the relationship between these features and a specific outcome, as well as the connection between the features themselves. The goal of this type of technology is not to tell the user what changes in his results specifically but to allow him to get a feel from an arbitrary starting point.

Where models are simple enough, these approaches seem to achieve the educational goals of both ECOA and the GDPR by allowing data subjects to gain an intuitive feel for the system. Ironically, this would be accomplished by complying with neither law because a person will not know a specific reason for denial or have an account of a model’s logic after playing with it, even if they feel that they understand the model better afterward.

While regulators have expressed interest in this idea, however, it poses a technical challenge. The statistical relationships at work in these models may be sufficiently complex that no consistent rule may become evident by tinkering with adjustable sliders. Models might involve a very large number of inputs with complex and shifting interdependencies such that even the most systematic tinkering would generate outcomes that would be difficult for a person to explain in a principled way.

One danger of this approach, then, is that it could do more to placate than elucidate. People could try to make sense of variations in the observed outputs by favoring the simplest possible explanation that accounts for the limited set of examples generated by playing with the system. Such an explanation is likely to take the form of a rule that incorrectly assigns a small set of specific variables unique significance and treats their effect on the outcome as linear, monotonic, and independent. Thus, for already simple models that can be explained, interactive approaches may be useful for giving people a feel without disclosing the algorithm, but for truly inscrutable systems, they could well be dangerous.

---


208. See Hildebrandt & Koops, supra note 207, at 450.

209. See id.

Remarkably, the techniques available within machine learning for ensuring interpretability correspond well to the different types of explanation required by existing law. There are, on the one hand, varied strategies and techniques available to practitioners that can deliver models whose inner workings can be expressed succinctly and sensibly to a human observer, whether an expert (e.g., a regulator) or lay person (e.g., an affected consumer). Laws like the GDPR that seek logic-like explanations would be well served by these methods. On the other hand, outcome-focused laws like ECOA that care only about principal reasons—and not the set of rules that govern all decisions—have an obvious partner in tools that furnish post hoc accounts of the factors that influenced any particular determination.

Where they succeed, these methods can be used to meet the demands of regulatory regimes that demand outcome- and logic-like explanations. Both techniques have their limitations, however. If highly sophisticated machine learning tools continue to be used, interpretability may be difficult to achieve in some instances, especially when the phenomena at issue are themselves complex. Post hoc accounts that list the factors most relevant to a specific decision may not work well when the number of relevant factors grows beyond a handful—a situation that is most likely to occur when such methods would be most attractive.

Notably, neither the techniques nor the laws go beyond describing the operation of the model. Though they may help to explain why a decision was reached or how decisions are made, they cannot address why decisions happen to be made that way. As a result, standard approaches to explanation might not help determine whether the particular way of making decisions is normatively justified.

III. FROM EXPLANATION TO INTUITION

So far, the majority of discourse around understanding machine learning models has seen the proper task as opening the black box and explaining what is inside. Where Part II.A discussed legal requirements and Part II.B discussed technical approaches, here we discuss the motivations for both. Based on a review of the literature, scholars, technologists, and policymakers seem to have three different beliefs about the value of opening the black box. The first is a fundamental question of autonomy, dignity, and

211. See supra note 16 and accompanying text.

212. These three rationales seem to track the rationales for ECOA’s adverse action notices as described in Part II.A.1. There is also scholarship that offers a fourth rationale, which includes due process and rule-of-law concerns. We set these concerns aside because they pertain to government use of algorithms, while this Article focuses on regulation of the private sector. See Brennan-Marquez, supra note 19, at 1288–94 (discussing “rule-of-law” principles with respect to police and judicial actions); Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 GEO. L.J. 1147, 1184–90, 1206–09 (2017) (discussing due process and reason-giving in administrative law); ECLT Seminars, [HUMIL16] 03: Katherine Strandburg, Decision-Making, Machine Learning and the Value of Explanation, YOUTUBE (Jan. 23, 2017), https://www.youtube.com/
personhood. The second is a more instrumental value: educating the subjects of automated decisions about how to achieve different results. The third is a more normative question—the idea that explaining the model will allow people to debate whether the model’s rules are justifiable.

The black-box-only approach is limited for the purposes of justifying decision-making. The first two beliefs are not about justifying decisions at all, and therefore serve a different purpose. The third is explicitly about justification, so our critique is directed not at its intent, but its operation. For those concerned with the justification for decision-making, the goal of explanation should be to find a way to bring intuition to bear in deciding whether the model is well justified. This Part explains both the power and limitations of such an approach.

A. The Value of Opening the Black Box

This Part identifies and elaborates the three rationales that apparently underlie most of the popular and scholarly calls for explanation.

1. Explanation as Inherent Good

There are several reasons to view explanation as a good unto itself, and perhaps a necessary part of a system constrained by law, including a respect for autonomy, dignity, and personhood. There is a fundamental difference between wanting an explanation for its own sake and wanting an explanation for the purpose of vindicating certain specific empowerment or accountability goals. Fears about a system that lacks explanation are visceral. This fear is best exemplified in popular consciousness by Franz Kafka’s The Trial, a story about a faceless bureaucracy that makes consequential decisions without input or understanding from those affected.

This concern certainly motivates some lawmakers and scholars. In his article, “Privacy and Power,” Daniel Solove refers to this as a “dehumanizing” state of affairs characterized by the “powerlessness and vulnerability created by people’s lack of any meaningful form of participation” in the decision. David Luban, Alan Strudler, and David Wasserman argue that “one central aspect of the common good”—which they argue forms the basis of law’s legitimacy—“lies in what we might call the moral intelligibility of our lives” and that the “horror of the bureaucratic process lies not in officials’ mechanical adherence to duty, but rather in the

watch?v=LQj3nbfSkrU (discussing procedural due process and explanations).


216. Id. at 1423.
individual’s ignorance of what the fulfillment of his or her duty may entail.”217 The concerns of dignity and personhood certainly motivate the data protection regime in Europe,218 if less directly the law in the United States.219

We lack the space (and the expertise) to do proper justice to the personhood argument for explanation. Accordingly, our goal here is to flag it and set it aside as a concern parallel to our broader concerns about enabling justifications for automated decisions.

To the extent that the personhood rationale can be converted to a more actionable legal issue, it is reflected in the concept of “procedural justice,” which was most famously championed by Tom Tyler. Procedural justice is the essential quality of a legal system that shows respect for its participants, which might entail transparency, consistency, or even politeness.220 Tyler and others have shown that people care deeply about procedural justice, to the point that they might find a proceeding more tolerable and fair if their procedural-justice concerns are satisfied even if they do not obtain their preferred outcome in the proceeding.221 Procedural justice, Tyler argues, is necessary on a large scale because it allows people to buy into the legal system and voluntarily comply with the law, both of which are essential parts of a working and legitimate legal system.222 Presumably, to the extent that automated decisions can be legally or morally justified, people must accept them rather than have them imposed, and as a result, the personhood rationale for model explanation also implicates procedural justice.

Ultimately, that there is inherent value in explanation is clear. But as a practical matter, those concerns are difficult to administer, quantify, and compare to other concerns. Where there are genuine trade-offs between explanation and other normative values such as accuracy or fairness, the inherent value of explanation neither automatically trumps competing considerations nor provides much guidance as to the type of explanation required. Therefore, while inherent value cannot be ignored, other rationales remain important.

---

221. See, e.g., Tom R. Tyler, Procedural Justice, Legitimacy, and the Effective Rule of Law, 30 CRIME & JUST. 283, 291 (2003); Tyler, supra note 220, at 128.
2. Explanation as Enabling Action

For others, the purpose of explanation extends to providing actionable information about the rendering of decisions, such that affected parties can learn if and how they might achieve a different outcome. Explanations are valuable, on this account, because they empower people to effectively navigate the decision-making process. Such beliefs are evident in the adverse action notice requirements of credit-scoring regulations, but they have come to dominate more recent debates about the regulatory function of requiring explanations of model-driven decisions more generally.

Across a series of recent papers, the debate has coalesced around two distinct, but related, questions. The first is whether and when the GDPR requires explanations of the logic or outcome of decision-making. The second is how to best explain outcomes in an actionable way.

The first question, whether to focus on outcome- or logic-based explanations, originates with an article by Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. These scholars split explanations between “system functionality” and “specific decisions”—a distinction functionally similar to our outcome- and logic-based framework. This mirrors the debate in the technical community about the best way to understand the meaning of interpretability. As described in Part II.B, the main split is whether to aim for interpretable models or to account for specific decisions. Drawing together the legal and machine learning literature, Lilian Edwards and Michael Veale have created a similar, but slightly altered distinction between “model-centric” and “subject-centric” explanations. While not identical, subject-centric explanations are another way to explain specific outcomes to individuals.

As the discussion has evolved in both the legal and computer science scholarship, new work has converged on the belief that explaining specific outcomes is the right approach. The debate has therefore shifted to the

223. See supra Part II.A.1.
224. Wachter et al., supra note 23.
225. Id. at 78. As Wachter and colleagues define it, system functionality is “the logic, significance, envisaged consequences, and general functionality of an automated decision-making system,” and explanations of specific decisions are “the rationale, reasons, and individual circumstances of a specific automated decision.” Id. While the distinction is broadly useful, our definitions differ from theirs and we believe the line between outcome- and logic-based explanations is less clear than they suggest. See Selbst & Powles, supra note 90, at 239 (arguing that, given the input data, a description of the logic will provide a data subject with the means to determine any particular outcome, and thus, explanations of the logic will often also explain individual outcomes).
226. Edwards & Veale, supra note 143, at 55–56. They define these terms as follows: “Model-centric explanations (MCEs) provide broad information about a [machine learning] model which is not decision or input-data specific,” while “[s]ubject-centric explanations (SCEs) are built on and around the basis of an input record.”
227. Ultimately, Edwards and Veale argue, as we do, that the explanation debate had been restricted to this question. Id. Recognizing that explanations are no panacea, the rest of their paper argues that the GDPR provides tools other than a right to explanation that could be more useful for algorithmic accountability.
second question, which focuses on the many different methods by which outcomes can be explained.

An interdisciplinary working group at the Berkman Klein Center for Internet and Society begin by recognizing that explanations are infinitely variable in concept, but claim that “[w]hen we talk about an explanation for a decision, . . . we generally mean the reasons or justifications for that particular outcome, rather than a description of the decision-making process in general.”228 They propose three ways to examine a specific decision: (1) the main factors in a decision, (2) the minimum change required to switch the outcome of a decision, and (3) the explanations for similar cases with divergent outcomes or divergent cases with similar outcomes.229 Wachter, Mittelstadt, and Chris Russell have a still narrower focus, writing about counterfactual explanations that represent “the smallest change to the world” that would result in a different answer.230 They envision a distance metric where, if one were to plot all \( n \) features in an \( n \)-dimensional space, the counterfactual is the shortest “distance” from the data subject’s point in the space (defined by the values of the features she possesses) to the surface that makes up the outer edge of a desirable outcome.231 Accordingly, counterfactual explanations are seen as fulfilling the three goals of explanations discussed in this Part: (1) to help an individual understand a decision, (2) to enable that individual to take steps to achieve a better outcome, and (3) to provide a basis for contesting the decision.232 When applying the strategy of counterfactual explanations, however, it is clear that most of the value comes from the second rationale: actionable explanations. Wachter and colleagues assert that counterfactual explanations are an improvement over the existing requirements of the GDPR because, as a matter of positive law, the Regulation requires almost nothing except a “meaningful overview,” which can be encapsulated via pictorial “icons” depicting the type of data processing in question.233 Counterfactual explanations, in contrast, offer something specific to the data subject and will thus be more useful in informing an effective response. But if their interpretation of the law is correct—that the GDPR requires no

---

228. Doshi-Velez & Kortz, supra note 201, at 2.
229. Id. at 3.
230. Wachter et al., supra note 143, at 845.
231. Id. at 850–54. Distance metrics are a way to solve this problem. Hall and colleagues describe another distance metric that is used in practice. Hall et al., supra note 120. They employ a distance metric to identify the features that need to change the most to turn a credit applicant into the ideal applicant. Id. Alternatively, other methods could be identifying the features over which a consumer has the most control, the features that would cost a consumer the least to change, or the features least coupled to other life outcomes and thus easier to isolate. The main point is that the law provides no formal guidance as to the proper metric for determining what reasons are most salient, and this part of the debate attempts to resolve this question. See 12 C.F.R. § 1002.9 supp. I (2018).
232. Wachter et al., supra note 143, at 843.
233. Id. at 865.
explanation—then their claim is that counterfactuals offer more than literally nothing, which is not saying much. On contestability, Wachter, Mittelstadt, and Russell ultimately concede that to contest a decision, it is likely necessary to understand the logic of decision-making rather than to just have a counterfactual explanation of a specific decision. The real value, then, of their intervention and others like it, is to better allow data subjects to alter their behavior when a counterfactual suggests that a decision is based on alterable characteristics.

Empowering people to navigate the algorithms that affect their lives is an important goal and has genuine value. This is a pragmatic response to a difficult problem, but it casts the goal of explanations as something quite limited: ensuring people know the rules of the game so they can play it better. This approach is not oriented around asking if the basis of decisions is well justified; rather it takes decisions as a given and seeks to allow those affected by them to avoid or work around bad outcomes. Rather than using explanations to ask about the justifications for decision-making, this approach shifts responsibility for bad outcomes from the designers of automated decisions to those affected by them.

3. Explanation as Exposing a Basis for Evaluation

The final value ascribed to explanation is that it forces the basis of decision-making into the open and thus provides a way to question the validity and justifiability of making decisions on these grounds. As Pauline Kim has observed:

234. The positive law debate about the right to explanation is not the subject of this Article, but suffice it to say, there is a healthy debate about it in the literature. See supra note 143 and accompanying text for a discussion.

235. Wachter et al., supra note 143, at 878. Their one example where a counterfactual can lead to the ability to contest a decision is based on data being inaccurate or missing rather than based on the inferences made. Thus, it is actually the rare situation specifically envisioned by FCRA, where the adverse action notice reveals that a decision took inaccurate information into account. Because of the deficiencies of the FCRA approach, discussed supra in Part II.A, this will not solve the general problem.

236. As Berk Ustun and colleagues point out, an explanation generated by counterfactual techniques will not necessarily be actionable unless intentionally structured to be so. Berk Ustun et al., Actionable Recourse in Linear Classification 2 (Sept. 18, 2018) (unpublished manuscript), https://arxiv.org/abs/1809.06514 [https://perma.cc/RPJ4-P4AP].

237. Mireille Hildebrandt, Primitives of Legal Protection in the Era of Data-Driven Platforms, 2 GEO. L. TECH. REV. 252, 271 (2018) (“Though it is important that decisions of automated systems can be explained (whether ex ante or ex post; whether individually or at a generic level), we must keep in mind that in the end what counts is whether such decisions can be justified.”).

238. This is remarkably similar to the longstanding privacy and data protection debate around notice and consent, where the goal of notice is to better inform consumers and data subjects, and the assumption is that better information will lead to preferable results. See generally Daniel J. Solove, Privacy Self-Management and the Consent Dilemma, 126 HARV. L. REV. 1880 (2013). In reality, this often fails to protect privacy because it construes privacy as a matter of individual decision-making that a person can choose to protect rather than something that can be affected by others with more power. See, e.g., Roger Ford, Unilateral Invasions of Privacy, 91 NOTRE DAME L. REV 1075 (2016).
When a model is interpretable, debate may ensue over whether its use is justified, but it is at least possible to have a conversation about whether relying on the behaviors or attributes that drive the outcomes is normatively acceptable. When a model is not interpretable, however, it is not even possible to have the conversation.239

But what does it mean to have a conversation based on what an interpretable model reveals?

In a seminal study, Rich Caruana and colleagues provide an answer to that question.240 They discovered that a model trained to predict complications from pneumonia had learned to associate asthma with a reduced risk of death.241 To anyone with a passing knowledge of asthma and pneumonia, this result was obviously wrong. The model was trained on clinical data from past pneumonia patients, and it turns out that patients who suffer from asthma truly did end up with better outcomes.242 What the model missed was that these patients regularly monitored their breathing, causing them to go to the hospital earlier.243 Then, once at the hospital, they were considered higher risk, so they received more immediate and focused treatment.244 Caruana and colleagues drew a general lesson from this experience: to avoid learning artifacts in the data, the model should be sufficiently simple that experts can inspect the relationships uncovered to determine if they correspond with domain knowledge. Thus, on this account, the purpose of explanation is to permit experts to check the model against their intuition.

This approach assumes that when a model is made intelligible, experts can assess whether the relationships uncovered by the model seem appropriate, given their background knowledge of the phenomenon being modeled. This was indeed the case for asthma, but this is not the general case. Often, rather than assigning significance to features in a way that is obviously right or wrong, a model will uncover a relationship that is simply perceived as strange. For example, if the hospital’s data did not reveal a dependence on an asthma diagnosis—which is clearly linked to pneumonia through breathing—but rather revealed a dependence on skin cancer, it would be less obvious what to make of that fact. It would be wrong to simply dismiss it as an artifact of the data, but it also does not fit with any intuitive story even a domain expert could tell.

Another example of this view of explanation is the approach to interpretability known as Local Interpretable Model-Agnostic Explanations (“LIME”).245 It has generated one of the canonical examples of the value of

239. Kim, supra note 4, at 922–23.
241. Id.
242. Id.
243. Id.
244. Id.
245. Ribeiro et al., supra note 197. This is one of the methods described supra in Part II.B.2.
interpretability in machine learning. Marco Ribeiro and colleagues used LIME to investigate a deep-learning model trained to distinguish images of wolves from huskies. The authors discovered that the model did not rely primarily on the animals’ features, but on whether snow appeared in the background of a photo.246

There are three reasons this is such a compelling example. First, what LIME identified as the distinguishing feature—snow—is legible to humans. Second, this feature is obviously not a property of the category “wolf.” Third, humans can tell a story about why this mistake occurred: wolves are more likely to be found in an environment with snow on the ground. Although this story may not actually be true, the important point is that we can convince ourselves it is.247 Like the asthma example, the ability to determine that the model has overfit the training data relies on the inherent legibility of the relevant feature, the existence of background knowledge about that feature, and our ability to use the background knowledge to tell a story about why the feature is important. In this example, the realization relies on something closer to common sense than to specialized expertise, but the explanation serves the same function—to allow observers to bring their intuition to bear in evaluating the model.

The final examples come from James Grimmelmann and Daniel Westreich,248 as well as Kim, whose work was discussed earlier.249 Grimmelmann and Westreich imagine a scenario in which a model learns to distinguish between job applicants on the basis of a feature—musical taste—that is both correlated with job performance and membership in a protected class.250 They further stipulate that job performance varies by class membership.251 As they see it, this poses the challenge of determining whether the model, by relying on musical tastes, is in fact relying on protected-class membership.252

Grimmelmann and Westreich then argue that if one cannot tell a story about why musical taste correlates with job performance, the model must be learning something else.253 They propose a default rule that the “something else” be considered membership in a protected class unless it can be shown

246. Ribeiro et al., supra note 197, at 1142–43. This is a textbook example of overfitting the training data.
247. In fact, while writing this section, we remembered the finding, but until we consulted the original source we disagreed with each other about whether the wolves or huskies were the ones pictured in snow. This suggests that the story would have been equally compelling if the error had been reversed.
248. Grimmelmann & Westreich, supra note 75.
249. Kim, supra note 4.
251. Id. at 167.
252. The only reason a model would learn to do this is if: (1) class membership accounts for all the variance in the outcome of interest or (2) class membership accounts for more of the variance than the input features. In the second case, the easy fix would be to include a richer set of features until class membership no longer communicates any useful information. The only way that adding features could have this effect, though, is if the original model was necessarily less than perfectly accurate, in which case a better model should have been used.
otherwise, specifically by the defendant. The problem with this reasoning is that the model might not be learning protected-class membership, but a different latent variable that explains the relationship between musical taste and job performance—an unobserved or unknown characteristic that affects both musical taste and job performance. By assuming that it should be possible to tell a story about such a variable if it exists, they—as in the examples above—fail to account for the possibility of a strange, but legitimate, result. They use the ability to tell a story as a proxy for the legitimacy of the decision-making, but that only works if a justification, or lack thereof, immediately falls out of the description, as it did in the asthma and snow examples.

Kim uses a real example to make a similar point. She cites a study stating that employees who installed web browsers that did not come with their computers stay longer on their job. She then speculates that either there is an unobserved variable that would explain the relationship or it is “entirely coincidental.” To Kim, what determines whether the relationship is “substantively meaningful” rather than a mere statistical coincidence is whether we can successfully tell ourselves such stories. Like Grimmelmann and Westreich, for Kim, if no such story can be told, and the model has a disparate impact, it should be illegal. What these examples demonstrate is that, whether one seeks to adjudicate model validity or normative justifications, intuition actually plays the same role.

Unlike the first two values of explanation, this approach has the ultimate goal of evaluating whether the basis of decision-making is well justified. It does not, however, ask the question: “Why are these the rules?” Instead, it makes two moves. The first two examples answered the question, “What are the rules?” and expected that intuition will furnish an answer for both why the rules are what they are and whether they are justified. The latter two examples instead argued that decisions should be legally restricted to intuitive relationships. Such a restriction short-circuits the need to ask why the rules are what they are by guaranteeing up front that an answer will be available.

254. Id. at 173.
255. Kim, supra note 4, at 922.
256. Id. So too did the chief analytics officer in the company involved, in an interview. Joe Pinsker, People Who Use Firefox or Chrome Are Better Employees, ATLANTIC (Mar. 16, 2015), https://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ [https://perma.cc/3MYM-SXAO] (“I think that the fact that you took the time to install Firefox on your computer shows us something about you. It shows that you’re someone who is an informed consumer,” he told Freakonomics Radio. “You’ve made an active choice to do something that wasn’t default.”).
257. Kim, supra note 4, at 917.
258. Id.
259. This might also explain the frequent turn to causality as a solution. Restricting the model to causal relationships also short-circuits the need to ask the “why” question because the causal mechanism is the answer. Ironically, a causal model need not be intuitive, so it may not satisfy the same normative desires as intuition seems to. See supra note 78.
These two approaches are similar, but differ in the default rule they apply to strange cases. In the case of the two technical examples, the assumption is that obviously flawed relationships will present themselves and should be overruled; relationships for which there is no intuitive explanation may remain. The two legal examples, by contrast, are more conservative. They presume that obviously correct relationships will show themselves, so that everything else should be discarded by default, while allowing for the possibility of defeating such a presumption. Both are forced to rely on default rules to handle strange, but potentially legitimate, cases because the fundamental reliance on intuition does not give them tools to evaluate these cases.

B. Evaluating Intuition

Much of the anxiety around inscrutable models comes from the legal world’s demands for justifiable decision-making. That decisions based on machine learning reflect the particular patterns in the training data cannot be a sufficient explanation for why a decision is made the way it is. Evaluating whether some basis for decision-making is fair, for example, will require tools that go beyond standard technical tests of validity that would already have been applied to the model during its development.260 While the law gives these tests some credence, reliance on accuracy is not normatively adequate with respect to machine learning.261

For many, the presumed solution is requiring machine learning models to be intelligible.262 What the prior discussion demonstrates, though, is that this presumption works on a very specific line of reasoning that is based on the idea that with enough explanation, we can bring intuition to bear in evaluating decision-making. As Kim observes:

Even when a model is interpretable, its meaning may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated.263

---

260. Even among practitioners, the interest in interpretability stems from warranted suspicion of the power of validation; there are countless reasons why assessing the likely performance of a model against an out-of-sample test set will fail to accurately predict a model’s real-world performance. Yet even with these deep suspicions, practitioners still believe in validation as the primary method by which the use of models can and should be justified. See Hand, supra note 182, at 12–13. In contrast, the law has concerns that are broader than real-world performance, which demand very different justifications for the basis of decision-making encoded in machine learning models.

261. Barocas & Selbst, supra note 4, at 673 (“[T]he process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.”).

262. See Brennan-Marquez, supra note 19, at 1253; Grimmelmann & Westreich, supra note 75, at 173; Kim, supra note 4, at 921–22.

263. Kim, supra note 4, at 922.
Her response is to constrain the model to features that bear an intuitive relationship to the outcome.\(^{264}\)

This way of thinking originates in disparate impact doctrine, which—among several ways of describing the requirement—calls for an employment test to have a “manifest relationship” to future job performance.\(^{265}\) But there is a difference between a manifest relationship of a model to job performance and a manifest relationship of a particular feature to job performance. Models can be shown to have a manifest relationship to job performance if the target variable is manifestly related to job performance and the model is statistically valid. This is true even if none of the individual features are manifestly related.\(^{266}\) People who advocate for a nexus between features and the outcome are dissatisfied with a purely statistical test and want some other basis to subject a model to normative assessment. Models must be restricted to intuitive relationships, the logic goes, so that such a basis will exist.

Regulatory guidance evinces similar reasoning. In 2011, the Federal Reserve issued formal guidance on model risk management.\(^{267}\) The purpose of the document was to expand on prior guidance that was limited to model validation.\(^{268}\) The guidance notes that models “may be used incorrectly or inappropriately” and that banks need diverse methods to evaluate them beyond statistical validation.\(^{269}\) Among other recommendations discussed in Part IV, the guidance recommends “outcomes analysis,” which calls for “expert judgment to check the intuition behind the outcomes and confirm that the results make sense.”\(^{270}\)

In an advisory bulletin about new financial technology, the Federal Reserve Board recommended that individual features have a “nexus” with creditworthiness to avoid discriminating in violation of fair lending laws.\(^{271}\) In their view, a nexus enables a “careful analysis” about the features assigned

\(^{264}\) Id.; cf. Nick Seaver, *Algorithms as Culture*, Big Data & Soc’Y, July–Dec. 2017, at 6 (“To make something [accountable] means giving it qualities that make it legible to groups of people in specific contexts. An accountable algorithm is thus literally different from an unaccountable one—transparency changes the practices that constitute it. For some critics, this is precisely the point: the changes that transparency necessitates are changes that we want to have.”).


\(^{266}\) Id. at 708.


\(^{268}\) Id. at 2.

\(^{269}\) Id. at 4.

\(^{270}\) Id. at 13–14.

\(^{271}\) Evans, *supra* note 121, at 4.
significance in a model predicting creditworthiness.\textsuperscript{272} Here, intuitiveness is read into ECOA as a natural requirement of having to justify decision-making that generates a disparate impact via the “business-necessity” defense.\textsuperscript{273} The business-necessity defense asks whether the particular decision-making mechanism has a tight enough fit with the legitimate trait being predicted\textsuperscript{274} and whether there were equally effective but less discriminatory ways to accomplish the same task. With a model that lacks intuitive relationships, a plaintiff could argue that the model is indirectly—and thus poorly—measuring some latent and more sensible variable that should serve as the actual basis of decision-making. The Federal Reserve Board guidance suggests that one way to avoid an uncertain result in such litigation is to limit decision-making to features that bear an intuitive—and therefore justifiable—relationship to the outcome of interest. While it is not clear that relying on proxies for an unrecognized latent variable presents problems under current disparate impact doctrine,\textsuperscript{275} the guidance treats an intuition requirement as a prophylactic. This reasoning seems to underlie the recommendations of Kim as well as Grimmelmann and Westreich.

What should be clear by now is that intuition is the typical bridge from explanation to normative assessment. This can be a good thing. Intuition is powerful. It is a ready mechanism by which considerable knowledge can be brought to bear in evaluating machine learning models. Such models are myopic, having visibility into only the data upon which they were trained.\textsuperscript{276} Humans, in contrast, have a wealth of insights accumulated through a broad range of experiences, typically described as “common sense.” This knowledge allows us to immediately identify and discount patterns that violate our well-honed expectations and to recognize and affirm discoveries that align with experience. In fact, intuition is so powerful that humans cannot resist speculating about latent variables or causal mechanisms when confronted by unexplained phenomena.

Intuition can also take the form of domain expertise, which further strengthens the capacity to see where models may have gone awry. The social sciences have a long history of relying on face validity to determine whether a model is measuring what it purports to measure.\textsuperscript{277} A model that assigns significance to variables that seem facially irrelevant is given little credence or is subject to greater scrutiny. Such a practice might seem ad hoc, but questioning face validity is a fundamental part of the social-scientific

\textsuperscript{272} Id.
\textsuperscript{273} It is interesting that the demand for intuitiveness, on this account, comes not from the procedural requirements of the adverse action notices—the part of ECOA most obviously concerned with explanations—but from the substantive concerns of disparate impact doctrine.
\textsuperscript{274} See, e.g., Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 1010 (1988) (Blackmun, J., concurring) (explaining that a business-necessity defense must be carefully tailored to objective, relevant job qualifications).
\textsuperscript{275} See Barocas & Selbst, supra note 4, at 709–10 (discussing the problems with the “fix-the-model” approach to alternative practice claims).
\textsuperscript{277} See supra note 73 and accompanying text.
process. Crucially, intuition allows us to generate competing explanations that account for the observed facts and to debate their plausibility.278

Importantly, however, intuition has its downsides. Most immediately, it can be wrong. It can lead us to discount valid models because they are unexpected or unfamiliar, or to endorse false discoveries because they align with existing beliefs.279 Intuition encourages us to generate “just so” stories that appear to make good sense of the presented facts. Such stories may feel coherent but are actually unreliable. In fact, the rich literature on cognitive biases—including the “narrative fallacy”—is really an account of the dangers of intuition.280 While intuition is helpful for assessing evidently good and bad results, it is less useful when dealing with findings that do not comport with or even run counter to experience. The overriding power of intuition means that strange results will stand out, but intuition may not point in a productive direction for making these any more sensible.

This is a particularly pronounced problem in the case of machine learning, as its value lies largely in finding patterns that go well beyond human intuition. The problem in such cases is not only that machine learning models might depart from intuition, but that they might not even lend themselves to hypotheses about what accounts for the models’ discoveries. Parsimonious models lend themselves to more intuitive reasoning, but they have limits—a complex world may require complex models. In some cases, machine learning will have the power to detect the subtle patterns and intricate dependencies that can better account for reality.

If the interest in explanation stems from its intrinsic or pragmatic value, then addressing inscrutability is worthwhile for its own sake. But if we are interested in whether models are well justified, then addressing inscrutability only gets us part of the way. We should consider how else to justify models. We should think outside the black box and return to the question: Why are these the rules?

IV. DOCUMENTATION AS EXPLANATION

Limiting explanation of a model to its internal mechanics forces us to rely on intuition to guess at why the model’s rules are what they are. But what would it look like for regulation to directly seek an answer to that question? By now, it is well understood that data are human constructs281 and that subjective decisions pervade the modeling and decision-making process.282

279. Raymond S. Nickerson, Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, 2 REV. GEN. PSYCHOL. 175, 175 (1998).
280. See generally KAHNEMAN, supra note 77.
282. Barocas & Selbst, supra note 4, at 673; see also Seaver, supra note 264, at 5.
Explaining why the model works as it does requires accounting for these decisions.

Furnishing such answers will require process, documentation, and access to that documentation. This can be done in a public format, with impact assessments, or companies can do it privately, with access triggered on some basis, like discovery in litigation.

A. The Information Needed to Evaluate Models

When we seek to evaluate the justifications for decision-making that relies on a machine learning model, we are actually asking about the institutional and subjective process behind its development. The Federal Reserve Board guidance discussed in Part III.B moves in this direction by recommending documentation, but its approach appears to be mostly about validation—how to validate well, thoroughly, on an ongoing basis, and in preparation for a future legal challenge. Careful validation is essential and nontrivial, but it is also not enough. Normatively evaluating decision-making requires, at least, an understanding of: (1) the values and constraints that shape the conceptualization of the problem, (2) how these values and constraints inform the development of machine learning models and are ultimately reflected in them, and (3) how the outputs of models inform final decisions.

To illustrate how each of these components work, consider credit scoring. What are the values embedded in credit-scoring models and under what constraints do developers operate? Lenders could attempt to achieve different objectives with credit scoring at the outset: Credit scoring could aim to ensure that all credit is ultimately repaid, thus minimizing default. Lenders could use credit scoring to maximize profit. Lenders could also seek to find ways to offer credit specifically to otherwise overlooked applicants, as many firms engaged in alternative credit scoring seek to do. Each of these different goals reflects different core values, but other value judgments might be buried in the projects as well. For example, a creditor could be morally committed to offering credit as widely as possible, while for others that does not factor into the decision. Or a creditor’s approach to regulation could be to either get away with as much as possible or steer far clear of regulatory scrutiny. Each of these subjective judgments will ultimately inform the way a project of credit scoring is conceived.

The developers of credit-scoring models will also face constraints and trade-offs. For example, there might be limits on available talent with both domain expertise and the necessary technical skills to build models. Models might be better informed if there were much more data available, even though

283. Bo. of Governors of the Fed. Reserve Sys., supra note 267. The guidance wants developers to consider where the data comes from, whether it suffers from bias, whether the model is robust to new situations, whether due care has been taken with respect to potential limitations and outright faults with the model, and so on. Id. at 5–16; see also Edwards & Veale, supra note 143, at 55–56; Pauline T. Kim, Auditing Algorithms for Discrimination, 166 U. Pa. L. Rev. Online 189, 196 (2017).
there are practical challenges to collecting so much data. Ultimately, both trade-offs are issues of cost, but they include more practical realities as well, such as limitations on talent in the geographical area of the firm or privacy concerns that limit the collection of more data. How to deal with these trade-offs is a judgment call every firm will have to make.

Another cost-related trade-off is competition. Before credit scoring was popular, creditors used to work with borrowers over the lifetime of the loan to ensure repayment; credit scores first took hold in banks as a way to reduce the cost of this practice. Creditors today could return to that model, but it would likely involve offering higher interest rates across the board to account for increased operating costs, perhaps pushing such a firm out of the market. As a result, competition operates as a constraint that ultimately changes the decision process.

The values of and constraints faced by a firm will lead to certain choices about how to build and use models. As we have discussed in prior work, the subjective choices a developer makes include choosing target variables, collecting training data, labeling examples, and choosing features. Developers must also make choices about other parts of the process, such as how to treat outliers, how to partition their data for testing, what learning algorithms to choose, and how and how much to tune the model, among other things. The act of developing models is quite complex and involves many subjective decisions by the developers.

In the credit example, the values discussed above may manifest in the model in several ways. For example, consider the different project objectives discussed above. If a firm seeks to maximize profit, it may employ a model with a different target variable than a firm that seeks to minimize defaults. The target variable is often the outcome that the model developers want to maximize or minimize, so in the profit-seeking case, it would be expected profit per applicant, and in the risk-based case, it could be likelihood of default. While the alternative credit-scoring model hypothesized above might rely on the same likelihood-of-default target variable, firms’ values are likely to influence the type of data they collect; they might seek alternative data sources, for example, because they are trying to reach underserved populations. In addition to the values embedded a priori, the values of the firms dictate how they resolve the different constraints they face—for example, cost and competition. The traditional credit scorers tend to not

289. Lehr & Ohm, supra note 51, at 683–700; see also Brian d’Alessandro, Cathy O’Neil & Tom LaGatta, Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification, 5 Big Data 120, 125 (2017).
make the extra effort or spend the extra money to obtain the data needed to make predictions about people on the margins of society. There is also regulatory uncertainty regarding the permissibility of new types of credit data. Therefore, their models reflect the fact that the developers are more sensitive to cost and regulatory penalty than inclusion.

Models are not self-executing; an additional layer of decisions concerns the institutional process that surrounds the model. Are the model outputs automatically accepted as the ultimate decisions? If not, how central is the model to the decision? How do decision makers integrate the model into their larger decision frameworks? How are they trained to do so? What role does discretion play?

These questions are all external to the model, but they directly impact the model’s importance and normative valence. For example, certain creditors may automatically reject applicants with a predicted likelihood of default that exceeds 50 percent. Others, however, may opt to be more inclusive. Perhaps a local credit union that is more familiar with its members and has a community-service mission might decide that human review is necessary for applicants whose likelihood of default sits between 40 percent and 60 percent, leaving the final decision to individual loan officers. A similar creditor might adopt a policy where applicants that the model is not able to score with confidence are subject to human review, especially where the outcome would otherwise be an automatic rejection of members of legally protected classes.

Many of these high-level questions about justifying models or particular uses of models are not about models at all, but whether certain policies are

291. Id. at 11,187–88.
292. The distinction between models and ultimate decisions is the focus of the GDPR’s prohibition on “decision[s] based solely on automated processing.” Article 29 Data Protection Working Party, supra note 149, at 19–22 (emphasis added).
293. This is not how credit typically works in the real world, but for demonstrative purposes, we decided to work with a single hypothetical. In reality, the best examples of this divergence between model and use come from policing and criminal justice. For example, the predictive-policing measure in Chicago, known as the Strategic Subject List, was used to predict the 400 likeliest people in a year to be involved in violent crime. Monica Davey, Chicago Police Try to Predict Who May Shoot or Be Shot, N.Y. TIMES (May 23, 2016), http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html [https://perma.cc/TZ2T-NMEJ]. When Chicago sought funding for the initiative, the city premised it on the idea of providing increased social services to those 400 people, but in the end only targeted them for surveillance. DAVID ROBINSON & LOGAN KOEPKE, STUCK IN A PATTERN: EARLY EVIDENCE ON “PREDICTIVE POLICING” AND CIVIL RIGHTS 9 (2016). The fairness concerns are clearly different between those use cases. See Selbst, supra note 4, at 142–44. Similarly, COMPAS, the now-infamous recidivism risk score, was originally designed to figure out who would need greater access to social services upon reentry to reduce the likelihood of rearrest but is now commonly used to decide whom to detain pending trial. Julia Angwin et al., Machine Bias, Propublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/F9CK-Z995].
acceptable independent of whether they use machine learning. Questions about justifying a model are often just questions about policy in disguise. For example, a predatory lender could use the exact same prediction of default to find prime candidates in underserved communities and offer them higher interest rates than they might otherwise receive. This will create more profit because the underserved loan candidates will be more willing to pay a higher rate, but it is clearly predation: interest rates are not being used to offset risk, but to extract maximum profit from vulnerable consumers. Most importantly, that this practice is predatory can be judged with no reference to the credit-scoring model.

Evaluating models in a justificatory sense means comparing the reasoning behind the choices made by the developers against society’s broader normative priorities, as expressed in law and policy. In order to perform this evaluation, then, documentation about the decisions that lie behind and become part of models must exist and be made available for scrutiny. With an understanding of what that information looks like, the next section begins to explore how to ensure access.

B. Providing the Necessary Information

Assuming the documentation exists, there are numerous ways it can become open to scrutiny. For purposes of demonstration, two are discussed here, although many more are possible: (1) the possibility that documentation is made publicly available from the start and (2) that it becomes accessible upon some trigger, like litigation. The former is essentially an algorithmic impact statement (AIS), a proposed variant of the original impact statements required by the National Environmental Policy Act. The most common trigger of the latter is a lawsuit, in which documents can be obtained and scrutinized and witnesses can be deposed or examined on the stand, but auditing requirements are another possibility. In both approaches, the coupling of existing documentation with a way to access it create answers to the question of what happened in the design process, with the goal of allowing overseers to determine whether those choices were justifiable. Like FCRA and ECOA, these examples have no inherent

294. See VIRGINIA EUBANKS, AUTOMATING INEQUALITY 37 (2018) (“[W]hen we focus on programs specifically targeted at poor and working-class people, the new regime of data analytics is more evolution than revolution. It is simply an expansion and continuation of moralistic and punitive poverty management strategies that have been with us since the 1820s.”).


296. According to sociologist Jacob Faber, this is actually what happened in the subprime crisis to people of color. Jacob W. Faber, Racial Dynamics of Subprime Mortgage Lending at the Peak, 28 HOUSING POL’Y DEBATE 328, 343 (2013).

297. Selbst, supra note 4, at 169–93.

connection to machine learning, but the methods can be easily applied in this context.

An impact statement is a document designed to explain the process of decision-making and the anticipated effects of that decision in such a way as to open the process up to the public. Generally, the requirement is designed to ensure that developers do their homework, create a public record, and include public comments.\textsuperscript{299} Impact statements are an idea that originated in 1970 with the National Environmental Policy Act \textsuperscript{300} and have since been emulated repeatedly at all levels of government, in many substantive areas of policy.\textsuperscript{301} Aside from environmental law, the federal government requires privacy impact assessments “when developing or procuring information technology systems that include personally identifiable information.”\textsuperscript{302} Individual states not only have their own legislation requiring environmental impact statements,\textsuperscript{303} but also racial impact statements for sentencing policy, among other requirements.\textsuperscript{304} Recently, led by the ACLU’s Community Control Over Police Surveillance (CCOPS) initiative,\textsuperscript{305} counties and cities have begun requiring impact statements that apply to police purchases of new technology.\textsuperscript{306}

One of us has argued that a future AIS requirement should be expressly modeled on the environmental impact statement (EIS): the original and most thorough version, with the fullest explanation requirements. Such an impact statement would require thoroughly explaining the types of choices discussed above. This includes direct choices about the model, such as target variables, whether and how new data was collected, and what features were considered. It also requires a discussion of the options that were considered but not chosen, and the reasons for both.\textsuperscript{307} Those reasons would—either explicitly or implicitly—include discussion of the practical constraints faced by the developers and the values that drove decisions. The AIS must also discuss the predicted impacts of both the chosen and unchosen paths, including the

\textsuperscript{299} Selbst, \textit{supra} note 4, at 169.
\textsuperscript{300} 42 U.S.C. §§ 4321–4347.
\textsuperscript{303} \textit{E.g.}, California Environmental Quality Act (CEQA), CAL. PUB. RES. CODE §§ 21000–21178 (2018).
\textsuperscript{304} Jessica Erickson, Comment, \textit{Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System}, 89 WASH. L. REV. 1425, 1445 (2014).
\textsuperscript{306} \textit{See, e.g.}, SANTA CLARA COUNTY, CAL., \textit{CODE OF ORDINANCES} § A40-3 (2016).
\textsuperscript{307} Selbst, \textit{supra} note 4, at 172–75.
possibility of no action, and the effects of any potential mitigation procedures.\footnote{308}

The typical American example of an impact statement is a public document. Thus, a law requiring them would also require that the developers publish the document and allow for comments between the draft and final impact statements.\footnote{309} Of course, such an idea is more palatable in the case of regulation of public agencies. While disclosure of the kinds of information we describe does not actually imply disclosure of the model itself—obviating the need for a discussion of trade secrets and gaming—firms may still be reluctant to publish an AIS that reveals operating strategy, perceived constraints, or even embedded values. Thus, it is also useful to consider a documentation requirement that allows the prepared documents to remain private but available as needed for accountability.\footnote{310}

A provision of the GDPR actually does just this. Article 35 requires “data protection impact assessments” (DPIAs) whenever data processing “is likely to result in a high risk to the rights and freedoms of natural persons.”\footnote{311} As Edwards and Veale discuss, the DPIA requirement is very likely to apply to machine learning,\footnote{312} and the assessments require “appropriate technical and organizational measures” to protect data subject rights.\footnote{313} In Europe, DPIAs are private documents, though making summaries public is officially encouraged.\footnote{314} The European solution to making this private document available is to require consultation with the member state data protection authorities whenever the DPIA indicates a high risk of interference with data subject rights.\footnote{315}

One could imagine another way of making an essentially private impact assessment accessible, initiated by private litigation. Interrogatories, depositions, document subpoenas, and trial testimony are all tools that enable litigation parties to question human witnesses and examine documents. These are all chances to directly ask model developers what choices they made and why they made them.

A hypothetical will help clarify how these opportunities, coupled with documentation—whether a DPIA or something similar—differ from the use of intuition as a method of justification. Imagine a new alternative credit-scoring system that relies on social media data.\footnote{316} This model assigns

\footnote{308. \textit{Id.}}
\footnote{309. \textit{Id.} at 177.}
\footnote{311. GDPR, \textit{supra} note 12, art. 35.}
\footnote{312. Edwards & Veale, \textit{supra} note 143, at 77–78.}
\footnote{313. GDPR, \textit{supra} note 12, art. 35.}
\footnote{315. Edwards & Veale, \textit{supra} note 143, at 78.}
\footnote{316. \textit{See, e.g., Astra Taylor \\& Jathan Sadowski, How Companies Turn Your Facebook Activity into a Credit Score,} NATION (May 27, 2015), https://www.thenation.com/article/how-companies-turn-your-facebook-activity-credit-score/ [https://perma.cc/P9FW-DSTN].}
significance to data points that are unintuitive but reliably predict default. Suppose the model also evinces a disparate impact along racial lines, as revealed by investigative journalists.

Black applicants denied credit then bring suit under the substantive nondiscrimination provisions of ECOA. Assuming, reasonably, that the judge agrees that disparate impact is a viable theory under ECOA, the case will turn on the business-necessity defense. Thus, in order to determine whether there was a legal violation, it is necessary to know why the designer of the model proceeded in using the particular features from social media and whether there were equally effective alternatives with less disparate impact.

Under an intuition-driven regime, such as that proposed by either Kim or Grimmelmann and Westreich, the case would begin with a finding of prima facie disparate impact, and then, to evaluate the business-necessity defense, the plaintiffs might put the lead engineer on the stand. The attorney would ask why social media data was related to the ultimate judgment of creditworthiness. The engineer would respond that the model showed they were related: “the data says so.” She is not able to give a better answer because the social media data has no intuitive link to creditworthiness. Under their proposed regime, the inquiry would end. The defendant has not satisfied its burden and would be held liable.

Under a regime of mandated documentation and looking beyond the logic of the model, other explanations could be used in the model’s defense. Rather than be required to intuitively link the social media data to the creditworthiness, the engineer would be permitted to answer why the model relies on the social media data in the first place. The documentation might show, or the engineer might testify, that her team tested the model with and without the social media data and found that using the data reduced the disproportionate impact of the model. Alternatively, the documentation might demonstrate that the team considered more intuitive features that guaranteed similar model performance but discovered that such features were exceedingly difficult or costly to measure. The company then used social media data because it improved performance and reduced disparate impact under the practical constraints faced by the company.


318. The engineer might have been able to come up with a story for why social media relates to credit—perhaps many of the applicant’s friends have low credit scores and the operating theory is that people associate with others who have similar qualities—and under this regime, such a story might have satisfied the defense. But the engineer knows this is a post hoc explanation that may bear little relationship to the actual dynamic that explains the model.

319. Grimmelmann & Westreich, supra note 75, at 170.

These justifications are not self-evidently sufficient to approve of the credit model in this hypothetical. Certainly, reducing disparate impact seems like a worthwhile goal. In fact, prohibiting or discouraging decision makers from using unintuitive models that exhibit any disparate impact may have the perverse effect of maintaining a disparate impact. Cost is a more difficult normative line and would likely require a case-by-case analysis. While intuition-based evaluation—and its reliance on default rules—would forbid the consideration of either of these motivations for using social media data, both rationales should at least enter into the discussion.

Having to account for all the decisions made in the process of project inception and model development should reveal subjective judgments that can and should be evaluated. This kind of explanation is particularly useful where intuition fails. In most cases, these decisions would not be immediately readable from the model. Recall that intuition is most useful where explanations of a model reveal obviously good or bad reasons for decision-making but will often offer no help to evaluate a strange result. Documentation will help because it provides a different way of connecting the model to normative concerns. In cases where the individual features are not intuitively related to the outcome of interest but there is an obviously good or bad reason to use them anyway, documentation will reveal those reasons where explanation of the model will not. Accordingly, these high-level explanations are a necessary complement to any explanation of the internals of the model.

Documentation will not, however, solve every problem. Even with documentation, some models will both defy intuition and resist normative clarity. Regardless, a regime of documentation leaves open the possibility of developing other ways of asking whether this was a well-executed project, including future understanding of what constitutes best practice. As common flaws become known, checking for them becomes simply a matter of being responsible. A safe harbor or negligence-based oversight regime may emerge or become attractive as the types of choices faced by firms become known and standardized. Documentation of the decisions made will be necessary to developing such a regime.


322. Documentation provides a further benefit unrelated to explanation. If the requirement for an intuitive link is satisfied, then the case moves to the alternative practice prong, which looks to determine whether there was another model the creditor “refuses” to use. Cf. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012). Normally, a “fix-the-model” response will not be persuasive because it is difficult to tell exactly how it went wrong, and what alternatives the developers had. Barocas & Selbst, supra note 4, at 705. With documentation, the alternatives will be plainly visible because that is exactly what has been documented.

323. Barocas & Selbst, supra note 4, at 715.

While there will certainly still be strange results for which neither intuition nor documentation works today, the overall set of cases we cannot evaluate will shrink considerably with documentation available.

CONCLUSION

Daniel Kahneman has referred to the human mind as a “machine for jumping to conclusions.” Intuition is a basic component of human reasoning, and reasoning about the law is no different. It should therefore not be surprising that we are suspicious of strange relationships in models that admit no intuitive explanation at all. The natural inclination at this point is to regulate machine learning such that its outputs comport with intuition.

This has led to calls for regulation by explanation. Inscrutability is the property of machine learning models that is seen as the problem, and the target of the majority of proposed remedies. The legal and technical work addressing the problem of inscrutability has been motivated by different beliefs about the utility of explanations: inherent value, enabling action, and providing a way to evaluate the basis of decision-making. While the first two rationales may have their own merits, the law has more substantial and concrete concerns that must be addressed. Those who believe solving inscrutability provides a path to normative evaluation also fall short because they fail to recognize the role of intuition.

Solving inscrutability is a necessary step, but the limitations of intuition will prevent normative assessment in many cases. Where intuition fails, the task should be to find new ways to regulate machine learning so that it remains accountable. Otherwise, maintaining an affirmative requirement for intuitive relationships will potentially impede discoveries and opportunities that machine learning can offer, including those that would reduce bias and discrimination.

Just as restricting evaluation to intuition will be costly, so would abandoning it entirely. Intuition serves as an important check that cannot be provided by quantitative modes of validation. But while there will always be a role for intuition, we will not always be able to use it to bypass the question of why the rules are the rules. We need the developers to show their work.

Documentation can relate the subjective choices involved in applying machine learning to the normative goals of substantive law. Much of the discussion surrounding models implicates important policy discussions, but does so indirectly. Often, when models are employed to change a way of making decisions, too much focus is placed on the technology itself instead of the policy changes that either led to the adoption of the technology or were wrought by its adoption. Quite aside from correcting one failure mode of intuition, documentation has a separate worth in laying bare the kinds of value judgments that go into designing these systems and allowing society to engage in a clearer normative debate in the future.

325. Kahneman, supra note 77, at 185.
326. See generally Eubanks, supra note 294.
We cannot and should not abandon intuition. But only by recognizing the role intuition plays in our normative reasoning can we recognize that there are other ways. To complement intuition, we need to ask whether people have made reasonable judgments about competing values under their real-world constraints. Only humans can answer these questions.
FACEBOOK’S “OVERSIGHT BOARD:” MOVE FAST WITH STABLE INFRASTRUCTURE AND HUMILITY

evelyn douek*

ABSTRACT

Facebook’s proposed Oversight Board is one of the most ambitious constitution-making projects of the modern era. With pre-existing governance of tech platforms delegitimized in the ongoing “techlash,” this represents a pivotal moment when new constitutional forms can emerge that will shape the future of online discourse. For all the potential of the Facebook Oversight Board, there are many things it cannot be. It will not hear a sufficient proportion of Facebook’s content moderation cases to be a meaningful response to calls for greater due process in individual decisions. Nor will it be able to become a font of globally accepted speech norms for the worldwide platform. The true value that the Board can bring to Facebook’s content moderation ecosystem lies between these two extremes of individual error correction and the settlement of globally applicable speech rules. The institutional offering of the Board should focus on two primary, but more modest, functions. First, it can help highlight weaknesses in the policy formation process at Facebook, removing blockages (such as blind spots and inertia) in the “legislative process” leading to the formulation of its Community Standards. Second, by providing an independent forum for the discussion of disputed content moderation decisions, the Board can be an important forum for the public reasoning necessary for persons in a pluralistic community to come to accept the rules that govern them, even if they disagree with the substance of those rules. Understanding the institutional role of the Board in these terms provides useful insights into the institutional design that will best help it achieve these goals.

I. Introduction ........................................................................................................................................... 2
II. The FOB’s Purpose ............................................................................................................................... 4
   A. Facebook’s Approach to the Impossible Task of Content Moderation............................................. 6
   B. The Benefits of Voluntary Restraints ................................................................................................. 8
      1. Bestowing legitimacy and reassuring users .................................................................................. 9
      2. Staving off or guiding regulation ............................................................................................... 12
      3. Outsourcing controversial decisions ......................................................................................... 13
      4. Enforcing existing standards .................................................................................................... 14
III. The FOB Blueprint ......................................................................................................................... 15
    A. Membership .................................................................................................................................... 16
    B. Power of Review ............................................................................................................................. 18
    C. Subject-Matter Jurisdiction ........................................................................................................... 20
IV. The Legitimacy Conundrum ........................................................................................................... 25
    A. The Limits of the FOB’s Legitimizing Power ................................................................................ 25
    B. The Limits of a “Constitution’s” Legitimizing Power .................................................................... 26
V. Maximizing the FOB’s Potential ........................................................................................................ 29
   A. Weakness as Strength ...................................................................................................................... 29
      1. The Benefits of a Judicial-Style Check on Policy-Making ......................................................... 30
      2. The Advantages of Weak-Form Review ..................................................................................... 31
      3. A Check on Users Too .............................................................................................................. 35
I. INTRODUCTION

Mark Zuckerberg, the Founder and Chief Executive Officer of Facebook, is engaged in one of the most ambitious constitutional projects of the modern era. On November 15, 2018, he announced in a blog post that by the end of 2019 Facebook will “create a new way for people to appeal content decisions to an independent body, whose decisions would be transparent and binding.”

Facebook has since released a “Draft Charter,” describing the “Oversight Board” as “a body of independent experts who will review Facebook’s most challenging content decisions – focusing on important and disputed cases.” In an earlier interview, when he first floated the idea, Zuckerberg analogized the proposed body to a “Supreme Court.”

Thus it seems that Zuckerberg is intending to introduce a check and balance into the governance of his sovereign domain of “Facebookistan.”

This innovation comes amidst ongoing constitutional upheaval for the internet. The “techlash” of the past few years, started by revelations of fake news and disinformation in the 2016 US election but sweeping broadly through the tech sector since then, has disrupted the status quo. Comparative scholarship shows that “[n]ew constitutional forms emerge only under extraordinary historical conditions, moments when pre-existing political, economic, and social structures have been weakened [or] deligitimized.”

This


6 Director of National Intelligence, Background to “Assessing Russian Activities and Intentions in Recent US Elections”: The Analytic Process and Cyber Incident Attribution Jan. 6, 2017.


8 ALEC STONE SWEET, GOVERNING WITH JUDGES: CONSTITUTIONAL POLITICS IN EUROPE 38 (2000).
moment represents such conditions for online platforms, and the changes that emerge will have significant ramifications for the future of online discourse. Because these online platforms mediate so much of modern life, sometimes called the “modern public square,” the effects will be far reaching.

Facebook is the most globally dominant social media company, with 2.32 billion monthly active users in countries around the world using the service in over 100 different languages. The public discourse that occurs on Facebook is central to political and cultural life in and between countries around the world. As companies everywhere are looking for solutions to the problems of content moderation, Facebook’s governance innovation could provide a model – or cautionary tale – for others. This is therefore a pivotal moment in the history of online speech governance.

For all the potential of the Facebook Oversight Board (hereafter, “FOB”), there are many things it cannot be. Politicians, academics, civil society and users have long been calling for greater transparency and due process in Facebook’s application of its Community Standards and other content moderation decisions. However, in many ways the FOB cannot be a meaningful answer to these calls. In other contexts, appeal or some form of “judicial review” of a decision can be a form of due process: such mechanisms give complainants an opportunity to voice their grievance, have a hearing, and receive some form of explanation for their treatment. Appeals processes can also be a way of ensuring the effective functioning of a bureaucratic system and rule enforcement by creating a mechanism for error correction. But the sheer volume of content moderation decisions Facebook makes every day means that the FOB cannot be expected to offer this kind of procedural recourse or error correction in more than the smallest fraction of these cases. This is compounded in Facebook’s case by the difficulty of ensuring the FOB’s decisions are effectively communicated and absorbed by the distributed and time-starved workforce of content moderators that make the first instance content moderation decisions.

Tellingly, due process or individual error correction are not the reasons Facebook says it is establishing the FOB. “Due process” is not mentioned in Zuckerberg’s blog post announcing the blueprint for the body or in the Draft Charter released a few months later. As for error correction, Zuckerberg says that artificial intelligence and more content moderators will help reduce mistakes but “we will never be perfect.” Facebook seems to conceive of the FOB as something different, and grander. In describing the

---

10 https://newsroom.fb.com/company-info/
15 See also Timothy Garton Ash et al., GLASNOST! Nine Ways Facebook Can Make Itself a Better Forum for Free Speech and Democracy 18–20 (2019).
17 See below at XX.
20 Zuckerberg, supra note 1.
function of the body, the Draft Charter explains that because decisions over what should and should not be allowed on Facebook are “too consequential for Facebook to make alone” the FOB will review the “most challenging content decisions” to provide oversight and make Facebook more accountable. It will be charged with reviewing enforcement of Facebook’s Community Standards – which “apply around the world to all types of content” in accordance with a set of indisputably worthy and sweeping values “like voice, safety, equity, dignity, equality and privacy.” However, if the vision of the FOB is to be a magisterial font of global speech norms for the global platform, this too is mistaken. The FOB will never be the ultimate arbiter of free speech norms whose pronouncements are accepted as legitimate in the way a court of final appeal might in a domestic legal system. It has neither the legitimacy nor authority necessary to fulfil this role.

The true value that the FOB can bring to Facebook’s content moderation ecosystem lies between these two extremes of individual error correction and the settlement of globally applicable speech rules. The institutional offering of the FOB should focus on two primary, but more modest, functions. First, it can help highlight weaknesses in the policy formation process at Facebook, removing blockages (such as blind spots and inertia) in the “legislative process” leading to the formulation of its Community Standards. Second, by providing an independent forum for the discussion of disputed content moderation decisions, the FOB can be an important forum for the public reasoning necessary for persons in a pluralistic community to come to accept the rules that govern them, even if they disagree with the substance of those rules. Understanding the institutional role of the FOB in these terms provides useful insights into the institutional design that will best help it achieve these goals.

This article proceeds in four parts. First, I look at the context for the introduction of the FOB into Facebook’s content moderation ecosystem. Content moderation – determining what should and shouldn’t be allowed on Facebook’s platform – is both an impossible and indispensable part of Facebook’s business. This paradox sheds light on why Facebook would voluntarily subject itself to the constraints of an independent appeal body. Key amongst the reasons for establishing the FOB is the desire to find a way of legitimizing the power that Facebook exercises over its users and the public sphere. It is only in understanding this context that we can begin to ask what would make the FOB’s institutional design effective. Second, I then sketch the current vision of the FOB based on early documents describing its institutional design. Third, I turn to the limitations on the FOB and the ways in which it differs from any prior institution that make the legitimizing goal especially difficult for the FOB. Finally, I turn to the ways in which Facebook can maximize the value the FOB can bring to its content moderation ecosystem in light of these inherent limitations.

The FOB is an experiment in governance that responds to an unprecedented amount of control by a private platform over the global public sphere. It will not and should not be a copy of any institution from a radically different context. This does not mean that there is not much that can be learnt from previous experience before the beta version of this updated form of platform governance is released.

II. THE FOB’S PURPOSE

---

22 Draft Charter at 1.
24 Draft Charter at 3.
26 JOHN RAWLS, POLITICAL LIBERALISM 212 (2005).
On the face of it, Zuckerberg and Facebook’s decision to set up the FOB might seem counterintuitive. Content moderation – platforms’ practice of designing and enforcing rules for what they allow to be posted on their services – is the commodity that platforms offer. The decision to move some of the power over this core part of Facebook’s business into independent hands may seem puzzling. Public pressure over Facebook’s decisions in recent years is not enough to explain the move. Zuckerberg is insulated from the need to respond to public pressure by his near absolute control of the company as founder, chief executive, chairman of the Board, and majority shareowner. But as Facebook’s Head of Policy Management has written: “[s]imply put, there are business reasons that a big social media company must pay attention to what the world thinks of its speech rules.” The goals of the FOB need to be understood in the context of how it might serve these business imperatives.

There is a much more familiar context in which holders of expansive unilateral power nevertheless employ a court system as a check and balance: rulers in authoritarian regimes. Scholars have previously drawn the comparison of Facebook as an autocracy due to Zuckerberg’s unchecked power over the Facebook “community.” The introduction of the FOB therefore raises the same question that comparative constitutional literature has grappled with: why do dictators allow courts any degree of judicial independence to courts that might interfere with their power? To be clear, I use the term “authoritarian” descriptively and without normative judgment. Facebook is a company, not a nation state. Authoritarianism is not unusual in this context. But the analogy is helpful in understanding why Facebook might introduce an “independent” check and balance its governance. There is a rich literature of the dynamics of courts in authoritarian regimes that helps explain how such institutions can further an autocrat’s aims.

After briefly sketching the impossible challenge of content moderation that the FOB is intended to help solve, this section draws on this literature to examine the reasons the FOB, and the renunciation of power that it represents, might appeal to Facebook as a solution.

---

30 David Pozen, Authoritarian Constitutionalism in Facebookland, Knight First Amendment Institute, https://knightcolumbia.org/content/authoritarian-constitutionalism-facebookland (drawing on Mark Tushnet’s characterization of absolutist constitutionalism as a system as one where “a single decisionmaker motivated by an interest in the nation’s well-being consults widely and protects civil liberties generally, but in the end, decides on a course of action in the decisionmaker’s sole discretion, unchecked by any other institutions.”); Balkin, supra note 12, at 2024; Henry Farrell O’Reilly Margaret Levi, and Tim, Mark Zuckerberg runs a nation-state, and he’s the king, Vox (Apr. 9–2018), https://www.vox.com/the-big-idea/2018/4/9/17214752/zuckerberg-facebook-power-regulation-data-privacy-control-political-theory-data-breach-king; MACKINNON, supra note 4; Chander, supra note 4.
32 Mehdi Shadmehr et al., Judicial Independence and Human Rights in Autocracies 40; RULE BY LAW: THE POLITICS OF COURTS IN AUTHORITARIAN REGIMES (Moustafa, Tamir & Ginsburg, Tom eds., 2008).
33 Andrew Keane Woods, Tech Firms Are Not Sovereigns, HOOVER INSTITUTION (2018).
A. Facebook’s Approach to the Impossible Task of Content Moderation

Consistent and coherent content moderation on the scale of a platform like Facebook is an essentially impossible challenge. Facebook has over 2 billion monthly active users,35 and over 2.5 billion pieces of content are shared on its platform per day.36 This content is moderated to align with Facebook’s public rules, called “Community Standards,”37 and its internal guidelines.38 In the first quarter of 2019 alone, Facebook “took action” on nearly 1.873 billion pieces of content for being in breach of these rules.39 Excluding the 1.8 billion pieces of content Facebook classified as spam, this means Facebook classified nearly 73 million pieces of content as falling within the other categories in its community standards: adult nudity and sexual activity, bullying and harassment, child nudity and sexual exploitation of children, hate speech, regulated goods, global terrorist propaganda and violence and graphic content. Because not every piece of content flagged for review is actually found to violate Facebook’s rules, these numbers represent only a subset of flagged items that Facebook evaluates.40 Facebook receives more than a million reports of violations of its content standards per day.41 As well as reviewing these reports, Facebook also re-reviews decisions that are appealed by users. In the first quarter of 2019, Facebook received nearly 25 million requests for appeal – around 275,000 requests per day.42 It is safe to say that this does not give the 30,000 people Facebook currently employs to enforce its Community Standards much time to consider each decision.43

Simply employing more people would not solve Facebook’s moderation woes (although it could certainly help). While a larger workforce might be able to more carefully consider each decision (for example, by looking at more contextual information that might shed light on the intended meaning of a particular post), this might come at the cost of greater consistency between decisions.44 Consistency is an important indicator of fairness. Because of this, and because of the sheer number of content moderation decisions Facebook has to make, the company has developed an “industrial” approach to content moderation.45 The goal is to create a “decision factory,” where application of Community Standards is reduced to bright-line rules, whose application is routinized and efficient.46 This attempted clean-cut approach has attracted controversy, such as over the strict use of “protected categories” in Facebook’s hate speech policy leading to attacks on “white men” contravening the Standards but not attacks on “black children.”47

Another tool Facebook uses to make content moderation at scale manageable is artificial intelligence

36 GILLESPIE, supra note 27, at 114.
37 Facebook, supra note 23.
39 Facebook, supra note 23.
40 Facebook releases the percentage of the pieces of content that it proactively identified using its own systems and those that were reported by users, but it does not identify how many reports it received for content that it determined were not violations of its Community Standards. See GILLESPIE, supra note 27, at 74.
41 Bickert, supra note 29, at 256.
43 Zuckerberg, supra note 1; For insight into the tough working conditions of front line content moderators, see SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).
45 GILLESPIE, supra note 27, at 77.
46 Caplan, supra note 44, at 24.
(AI). In his Blueprint, Zuckerberg calls AI “the single most important improvement in enforcing our policies,” because it can quickly and proactively identify harmful content. This further reflects Facebook’s mechanistic approach to content moderation. For every category except bullying and harassment and hate speech, Facebook found over 95% of the content it took down as violating its community standards before it was reported by a user, in large part because of its AI. But the exception of bullying and harassment and hate speech is telling: these two categories of content are harder for Facebook to proactively identify because they are so highly context-dependent. As Facebook says of bullying reports, these “tend to be personal and context-specific, so in many instances we need a person to report this behavior to us before we can identify or remove it. This results in a lower proactive detection rate than other types of violations.” Hate speech is notoriously difficult to detect through automated processes, because it depends so much on linguistic nuance, intention and local norms. As Ash et al. write, context is all important “at the complex frontiers of political speech, dangerous speech, and hate speech.” So while AI is a necessary part of content moderation at scale, it is not sufficient.

Even if it was technologically possible to train AI to appreciate the infinite spectrum of human nuance, there are at least two more reasons why AI cannot be complete answer to the content moderation problem.

First, AI does not give a person who has a decision made against them the substance or feeling of due process or of being heard. AI also does not give public reasoning for its decisions, which, as I argue below, is an essential component of a legitimate content moderation system.

The second reason AI cannot be a full answer to the difficulties of content moderation is even more fundamental—AI cannot determine the anterior question of the values that should be encoded into the detection algorithms. Before asking AI to identify and remove hate speech from the platform, for example, it has to be told what to look for. But what constitutes impermissible hate speech is an essentially contested concept that varies around the world. As one Facebook representative writes, “There is no universally accepted answer for when something crosses the line. Although a number of countries have laws against hate speech, their definitions of what constitute hate speech vary significantly.” A well-known example is Holocaust denial. As a fairly concrete category of speech, less context dependent than many

---

49 Zuckerberg, supra note 1.
51 Evelyn Douek, Zuckerberg’s New Hate Speech Plan: Out With the Court and In With the Code, Lawfare (Apr. 14, 2018), https://www.lawfareblog.com/zuckerbergs-new-hate-speech-plan-out-court-and-code (noting Zuckerberg acknowledges the difficulty of using AI to detect hate speech in Congressional testimony, and discussing why it is difficult); Zuckerberg also acknowledged as much in his Blueprint: Zuckerberg, supra note 1 (“As you get into hate speech and bullying, linguistic nuances get even harder … the state of the art in AI is still not sufficient to handle these challenges on its own. So we use computers for what they’re good at -- making basic judgements on large amounts of content quickly -- and we rely on people for making more complex and nuanced judgements that require deeper expertise.”).
52 Ash et al., supra note 15, at 11.
53 See generally Danielle Keats Citron, Technological Due Process, 85 WASH. UL REV. 1249 (2007).
54 See XX.
others, this might be a prime example of speech that AI might be able to more easily identify. But while Holocaust denial is illegal in several countries,\textsuperscript{57} it is famously not in the United States where the decision that Nazis should be allowed to march in front of Holocaust survivors is seen as one of the “truly great victories” in American legal history.\textsuperscript{58} AI itself cannot choose between these two opposing conceptions of free speech: it has to be told whether to find and remove Holocaust denial by a human. Another example is Facebook’s exception to its Community Standards for content it deems to be “newsworthy.”\textsuperscript{59} These cases require a balancing of the harm caused by allowing speech that breaches Facebook’s rules to remain on the platform with the public interest in being informed about the particular matter. Such balancing cannot be done in the abstract or ex ante and encoded into an algorithm, but requires case-by-case consideration.

Enter Facebook’s “Oversight Board,” which could temper the industrial application of content moderation rules at scale. The FOB can give a user an opportunity to be heard, and will then consider all the relevant context and competing values at stake in the case at hand. Most importantly, the FOB will offer an explanation for why content is or is not allowed on Facebook’s platform.

\textbf{B. The Benefits of Voluntary Restraints}

The question remains, however, why outsource this role to an independent body? Greater transparency and reason-giving could be provided by Facebook employees and policy-makers within the current content moderation ecosystem. And despite growing calls,\textsuperscript{60} Zuckerberg is showing no signs of wanting to relinquish any power at his company.\textsuperscript{61} His enjoyment of this power is a matter of legend.\textsuperscript{62}

Nevertheless, the FOB does constitute a renunciation of a degree of Facebook’s power. The FOB will have the power to reverse Facebook’s decisions,\textsuperscript{63} constraining the otherwise plenary discretion that Facebook currently has over what appears on its platform.\textsuperscript{64} Its establishment will recreate a system of separation of powers, where an independent “judicial”-style body will oversee the other branches of Facebook’s content moderation: the “legislative” branch that writes Facebook’s content moderation rules, and executive actors who implement these rules (with the help of AI). This is an unprecedented governance structure for a private company. But it is, of course, the dominant form of governance in

\begin{itemize}
\item \textsuperscript{58} Geoffrey R. Stone, Remembering the Nazis in Skokie, Huffington Post (May 20, 2009), https://www.huffingtonpost.com/geoffrey-r-stone/remembering-the-nazis-in_b_188739.html.
\item \textsuperscript{59} Kate Klonick, Facebook v. Sullivan, Knight First Amendment Institute Emerging Threats Series (2018).
\item \textsuperscript{60} Margaret Sullivan, Embattled and in over his head, Mark Zuckerberg should — at least — step down as Facebook chairman, Washington Post (Nov. 19, 2018), https://www.washingtonpost.com/lifestyle/style/embattled-and-in-over-his-head-mark-zuckerberg-should-at-least--step-down-as-facebook-chairman/2018/11/19/de00c2d8-ebfa-11e8-96d4-0d23f2aad09_story.html.
\item \textsuperscript{61} Kara Swisher, Full transcript: Facebook CEO Mark Zuckerberg on Recode Decode Recode, Recode (2018), https://www.recode.net/2018/7/18/17575158/mark-zuckerberg-facebook-interview-full-transcript-kara-swisher (confirming he has no intention of firing himself); Zuckerberg says stepping down at Facebook is “not the plan,” Engadget, (Nov. 20, 2018) https://www.engadget.com/2018/11/20/zuckerberg-says-stepping-down-at-facebook-is-not-the-plan/ (confirming he has no intention of stepping down as Chairman of the board.).
\item \textsuperscript{62} Zuckerberg famously had business cards that declared “I’m CEO, Bitch”: MG Siegler, Card Designer: The Inspiration For Zuckerberg’s “I’m CEO, Bitch”? Steve Jobs., TechCrunch (2010), http://social.techcrunch.com/2011/06/25/im-ceo-bitch/.
\item \textsuperscript{63} Draft Charter at 1.
\item \textsuperscript{64} Klonick, supra note 38, at 1617 (“These platforms are both the architecture for publishing new speech and the architects of the institutional design that governs it. Because of the wide immunity granted by § 230, these architects are free to choose which values they want to protect — or to protect no values at all.”).
\end{itemize}
nation states.

The matter should not be overstated – Zuckerberg is not recreating liberal democratic governance. He is not subjecting himself or his role to democratic accountability. But the FOB initiative is in keeping with Zuckerberg’s long-standing pronouncements that Facebook is “more like a government than a traditional company.”65 By initially referring to the FOB as a “Supreme Court” and calling its rules of operation a “Charter,” for example, Zuckerberg is implicitly distinguishing the FOB from other, more well-known forms of online dispute resolution (ODR) such as eBay’s Resolution Center.66 The message is that this is not ordinary commercial customer relations management; this is something grander. And the distinction is accurate. While most ODR systems are directed at resolving disputed between two private parties (such as a buyer and seller),67 the FOB will be resolving disputes more analogous to public law. The disputes it will hear will revolve around the exercise of power by the “government” of Facebook.68 In making decisions about the rules in what has become a key channel of public discourse, the FOB’s decisions will also need to take into account a broader conception of the “public interest,” rather than the mere resolution of the immediate dispute before it. So while many of these other ODR systems are increasingly looking to automate processes and displace human dispute handlers,69 the FOB looks to reintroduce and elevate the human component, recalling more conventional offline dispute resolution bodies.

There are four apparent reasons why this governance structure may appeal to Facebook to solving the content moderation dilemmas outlined above: (1) bestowing content-moderation decisions with an aura of legitimacy, aiding user-relations; (2) staving off or guiding more extensive governmental regulation; (3) outsourcing controversial decisions away from the company; and (4) facilitating better enforcement of existing standards. These closely track the reasons authoritarian regimes often retain court systems at least somewhat independent from the regime.70

1. Bestowing legitimacy and reassuring users

Legitimacy is central to the FOB experiment. I return to the topic of legitimacy at length below,71 but two points are worth noting here. First, “legitimacy” can mean many things and is hard to define.72 Throughout this paper, references to “legitimacy” are references to what Fallon calls “sociological

---

65 David Kirkpatrick, The Facebook Effect: The Inside Story of the Company That is Connecting the World 254 (2010) (quoting Zuckerberg as saying “In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we’re really setting policies.”)


68 Black’s Law Dictionary (10th ed. 2014) (Public law is defined as “[t]he body of law dealing with the relations between private individuals and the government, and with the structure and operation of the government itself”).

69 Katsh & Rabinovich-Einy, supra note 66, at 38.

70 Moustafa and Ginsburg suggest five functions for courts in an authoritarian state: “to (1) establish social control and sideline political opponents, (2) bolster a regime’s claim to “legal” legitimacy, (3) strengthen administrative compliance within the state’s own bureaucratic machinery and solve coordination problems among competing factions within the regime, (4) facilitate trade and investment, and (5) implement controversial policies so as to allow political distance from core elements of the regime.”; Tamir Moustafa & Tom Ginsburg, Introduction, in Rule by Law: The Politics of Courts in Authoritarian Regimes 4 (Tamir Moustafa & Tom Ginsburg eds., 2008).

71 See Part IV.

legitimacy” – that is, the prevailing public attitude towards Facebook’s content moderation. Importantly, legitimacy does not mean correctness; instead, “in circumstances of relatively widespread reasonable disagreement, . . . legitimacy connotes[ ]respect-worthiness.” Therefore, the mission to legitimate Facebook’s content moderation ecosystem aims to create a situation where its decisions are generally considered somewhat worthy of respect, even if there are those that still disagree with the substance of those decisions. The second thing worth noting is that this definition of legitimacy highlights the very low baseline from which Facebook is operating. In the past few years, Facebook’s decisions have increasingly come to be viewed as inconsistent and arbitrary, and therefore illegitimate. A recent Pew survey found that only 31% of US adults have a great deal or fair amount of confidence in social media companies to determine what offensive content should be removed from their platforms. For the FOB to be successful in conferring legitimacy to Facebook’s content moderation, it does not need to create a situation where content moderation comes to be viewed as perfectly legitimate; it only needs to improve over the current state of widespread perceptions of illegitimacy.

Independent court-like institutions are one tool for mitigating the extent to which an otherwise unconstrained ruler’s decisions are viewed as wholly illegitimate. As Dixon and Landau write, “[w]hen exercising powers of judicial review, most courts are . . . afforded a degree of presumptive legitimacy.” Authoritarian regimes use courts in this way: because they lack democratic procedural legitimacy in the exercise of power, such regimes often use courts to give a patina of more substantive legitimacy to their rule. Why should such regimes care about legitimacy in the first place? For authoritarian regimes, legitimacy is important because it can reduce the costs and necessity of resorting to force to maintain power. For Facebook the calculation is somewhat different but not entirely so. Because “code is law,” Facebook’s enforcement of its rules does not rely on force in the same way. However, such enforcement will have other costs if stakeholders oppose Facebook’s rulings. Facebook needs the approval of governments, users, advertisers, and the media in establishing its speech norms because these stakeholders can exert commercial pressure on the company—these are the “business reasons” that Facebook must pay attention to external reactions to how it polices its platform.

Indeed, because “exit” is easier than physical exit from a state, these costs may be even greater. While network effects make it more unlikely that Facebook will become the next Myspace, a social media graveyard of abandoned profiles, the last few years of scandals no doubt make Facebook afraid to be complacent. As a Facebook-commissioned report by a group of independent academics explained:

---

73 Id. at 21. (that is, this paper is not directly concerned with the legal or moral legitimacy of Facebook’s decisions, although these will of course affect sociological legitimacy.)

74 Id. at 8.


78 Moustafa, supra note 34, at 286.

79 Moustafa & Ginsburg, supra note 70, at 5.

80 LAWRENCE LESSIG, CODE V2.0 (Version 2.0 ed. 2006).


82 Bickert, supra note 29, at 265.

Facebook has considerable capacity to manage its content regulation process in a top down manner which pays minimal attention to users’ views. However, the very existence of the [Community Standards Enforcement Report] highlights the recognition that public views about Facebook and its attitude about the public matter. They matter for the individual user both because disgruntled users find ways to circumvent rules, for example opening multiple accounts. In addition, unhappy customers are less likely to use the site and more likely to seek alternatives to it.84

Put simply, authoritarians and companies value legitimacy because it can reduce the costs of decisions with which people disagree.85

Perceived legitimacy does not only improve Facebook’s relationship with its users, but also with its commercial partners such as advertisers. Such partners may be dissuaded from dealing with Facebook if it becomes too unpredictable or entails high reputational costs. Again, this may seem a distant threat given Facebook’s practical duopoly over online advertising, but it no doubt still factors into Facebook’s decisions. The FOB’s role in mitigating this threat again resembles the way authoritarian regimes use courts to respond to foreign investor fears over an unpredictable local business environment.86 As Moustafa observes, “[i]n the age of global competition for capital, it is difficult to find any government that is not engaged in some program of judicial reform designed to make legal institutions more effective, efficient, and predictable for the purpose of attracting global capital.”87 For example, Egypt bolstered its court processes in the late-1950s and 1960s to stem the severe capital flight it was experiencing because investors became wary of the consolidation of power in the Nasser regime without any credible legal institutional checks.88

For both users and commercial partners, then, an independent body overseeing content moderation can reassure these stakeholders, raising the value of the Facebook “product” of content moderation by making it appear more stable and less subject to Facebook’s whims.89

Courts are also a relatively low-cost way of achieving these legitimacy gains. For many authoritarian regimes, “[g]ranting access to the courts was a concrete way to relieve political pressure without opening the political system.”90 That is, a court-like check is a far more limited surrender of power, defined in scope, than wholesale reform of the policy-making institutions. Facebook may think that by surrendering some power over content moderation decisions, it is bolstering its claim to be running the community for the benefit of its users and thereby try to weaken calls for more extensive reforms to decision-making at the company.

87 Moustafa, supra note 34, at 285.
90 Moustafa, supra note 34, at 287.
2. Staving off or guiding regulation

Facebook may view the establishment of the FOB as a way of forestalling more extensive governmental regulation or at least as an opportunity to shape the form that such regulation may take. As tech platforms’ power and the related number of controversies grow, there is a growing sense that new laws are “inevitable.”91 Even Zuckerberg himself is now calling for global regulation of “harmful content.”92 In the absence of global regulation, an ad hoc patchwork of local regulations could be more costly for companies by requiring them to comply with multiple different sets of rules. Research shows that even relatively modest voluntary efforts by private firms to restrain their own behavior can stave off much more stringent public regulations.93 A good example of this is ad transparency on social media platforms. Facebook, Google, and Twitter have all recently unveiled voluntary ad transparency measures;94 meanwhile the proposed Honest Ads Act, which would compel these sorts of disclosures, has made little progress in Congress,95 despite extensive reporting about weaknesses in Facebook’s ad transparency tools.96 Another example is tech platforms entering into voluntary Code of Conduct agreements with the European Union, which held off coercive measures.97

Voluntary reforms have benefits for governments too. Effectively regulating tech companies will be extremely difficult – if done poorly, governments may be left taking the blame for poor outcomes. Self-regulatory reform at Facebook may make lack of governmental regulation more politically tenable for politicians by showing some form of progress. Authoritarian states use formal compliance with constitutional requirements to help maintain legitimacy. They recommended a model that focused on “expanding and legitimizing” platform self-regulation based on “the progress made in the last 12 months by … Facebook,”99 showing that self-imposed platform reform and regulatory reform occur in dialogue with one another.100

94 Issie Lapowsky & Louise Matsakis, You Can Now See All the Ads Facebook Is Running Globally, WIRED (06.28.18), https://www.wired.com/story/facebook-aims-more-transparency-view-ads-feature/.
97 Citron, supra note 12, at 1041–44.
98 Moustafa & Ginsburg, supra note 70, at 6 (giving the examples of postwar Korea and Taiwan, and the Philippines under Ferdinand Marcos).
99 French Secretary of State for Digital Affairs, Regulation of Social Networks – Facebook Experiment May 2019.
100 For more on this relationship between voluntary initiatives and government regulation, see Emily B. Laidlaw,
In some countries, the reform may be more anticipatory than others. In the US, the FOB might be seen as an attempt to preempt calls to narrow intermediary immunity provisions to deny platforms protection if they fail to take reasonable steps to prevent their services being used maliciously.¹⁰¹ In other countries, the reform may merely be responsive to certain regulations that have already been put in place. A notable example is recent regulation in Germany that Facebook’s head of policy for Europe has described as already “pushing us to the role of the court, the role of the legal system.”¹⁰² In either case though, Facebook’s establishment of the FOB makes it not merely a passive recipient of regulatory mandates but a proactive player in the design of the future of internet governance.

3. Outsourcing controversial decisions

For many of the issues that arise in the course of content moderation, there will be no “right” answer and any decision is likely to upset a certain constituency. The FOB allows Facebook to pass the responsibility for these divisive decisions to an independent body: a renunciation of power in the hope of also outsourcing some of the blame for contentious choices.

A desire for Facebook to no longer bear the brunt of public opprobrium for content moderation decisions is evident in Zuckerberg’s call for third-party bodies to set global standards for “harmful content.”¹⁰³ A more independent body set up by regulators would further distance Facebook from these choices, but global cooperation is unlikely to be forthcoming in the near future, if ever. Setting up the FOB might be seen as the next-best alternative. Democratic¹⁰⁴ and authoritarian regimes¹⁰⁵ alike use courts as a shield for controversy that can attend divisive political decisions.¹⁰⁶ Despite having “no influence over either the sword or the purse,”¹⁰⁷ courts are obeyed by other governmental actors at least in part because then “[d]isaffected citizens can then blame the [courts], not elected officials, for their disappointments.”¹⁰⁸

The motivation to avert blame is not limited to governments—other actors can also use the law to avoid social backlash for their decisions. For example, many voluntarily segregated restaurants and hotels lobbied for civil rights legislation that would have forced them to desegregate.¹⁰⁹ Segregation was

---

¹⁰³ Zuckerberg, supra note 92.
¹⁰⁵ Moustafa, supra note 34, at 286.
¹⁰⁶ Mark A. Graber, The Nonmajoritarian Difficulty: Legislative Deference to the Judiciary, 7 STUD. AM. POL. DEV. 35, 43 (1993)] (“the aim of legislative deference to the judiciary is for the courts to make controversial policies that political elites approve of but cannot publicly champion, and to do so in such a way that these elites are not held accountable by the general public, or at least not as accountable as they would be had they personally voted for that policy”); Vanberg, supra note 89, at 108 (“Rather than make decisions that carry an electoral risk, officials may prefer to pass such issues off to someone else, including the judiciary. However, this strategy of ‘passing the buck’ only works if the other actor is independent.”).
¹⁰⁷ The Federalist No. 78 (Alexander Hamilton).
¹⁰⁸ Fallon, supra note 72, at 115.
commercially costly for hotels, depriving them of black customers, but voluntarily desegregating would have imposed a social cost in their prejudiced communities. Civil rights legislation made desegregation a matter of legal compliance rather than a decision with social meaning, and outsourced the value judgment.

Similarly, when the FOB decides that a certain popular commentator should be banned for sprouting hate speech or that certain misleading political content should be left up rather than completely censored, Facebook can cast implementing these decisions as mere compliance rather than its own politically-charged decisions. This fits with Facebook’s own professed image as apolitical, and agnostic as to the ideology of content on its platform. Facebook does not want to take sides in the culture wars, presumedly because doing so would alienate segments of its customer base. After all, Republicans buy shoes from Facebook ads too.

Of course, there are limits to the extent to which Facebook can avoid responsibility for what it allows on its platform. As one representative has written, “To be clear, we are not asking a group of experts to make decisions for us. We are, however, asking for their insights to help inform our thinking and hold us accountable. We’ll still be making hard decisions every day, and we accept the full weight of that responsibility.” Facebook cannot be seen to be trying to wash its hands of important content decisions that define the user experience. Finding the right balance will no doubt involve a degree of ongoing calibration.

4. Enforcing existing standards

To a limited extent, the FOB might also help not only with changing or filling in gaps in Facebook’s policies but also with enforcing the policies that Facebook already has. In any complex system, enforcement error is inevitable. As Zuckerberg says in his Blueprint, “[t]he vast majority of mistakes we make are due to errors enforcing the nuances of our policies rather than disagreements about what those policies should actually be.” An appeals body can be a useful mechanism for monitoring the performance of and correcting errors in the application of its existing standards. It has the advantage of distributing the burden of finding mistakes by making private parties bear the cost of rectifying errors by bringing cases. This is how many authoritarian regimes use a system of administrative law courts, to resolve principal-agent problems in the administration of their policies that cannot resolved through centralized monitoring alone.

Facebook may, for example, be satisfied that it has drawn the appropriate line in its definition of hate speech. But this highly contextual and sensitive judgment call will not always be implemented correctly

---


113 Zuckerberg, supra note 1.


115 Facebook defines hate speech as “a direct attack on people based on what we call protected characteristics – race,
by context-blind AI tools or the burnt-out, time-pressured and geographically-distributed contractors that are charged with implementing these rules. Currently, it is not uncommon for external parties to find such errors and to extract apologies from Facebook once they draw public attention to the mistakes. Creating the FOB formalizes this process and redirects the public outrage into a predetermined process for the resolution of disputes. This has the added benefit of encouraging engagement with the platform and its rules, rather than mere criticism or flight. This process of engagement keeps users invested in the site and creates a kind of buy-in.

But it is important to be realistic about the extent of this function for the FOB: the FOB will only be able to hear and decide appeals in the tiniest fraction of content moderation decisions Facebook makes. Error correction cannot be its primary goal; it will be a collateral benefit only. And the extent to which regulators are persuaded to stay their hand and the public no longer blames Facebook alone for controversial decisions is dependent on whether the FOB itself is seen as legitimate and independent. Legitimacy is therefore the crux of the FOB experiment. Before discussing the fundamental challenge of creating legitimacy, the next section sketches the broad strokes of the FOB’s institutional design.

### III. THE FOB BLUEPRINT

Many details about the final institutional design of the FOB remain unknown at the time of writing. Zuckerberg’s Blueprint and the later released Draft Charter literally ask more questions than they answer. In his blog post, Zuckerberg asked: “how are members of the body selected? How do we ensure their independence from Facebook, but also their commitment to the principles they must uphold? How do people petition this body? How does the body pick which cases to hear from potentially millions of requests?” As one report put it, “[a]ll the major questions remain unanswered.” The Draft Charter starts to fill in some details on the blueprint, but itself acknowledges that it is only a “starting point.” Nevertheless, these documents do provide a rough picture of what the FOB will look like.

---

ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define “attack” as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.”: Facebook, supra note 23.


118 See SWEET, supra note 8, at 141 for a similar interpretation of the exercise of delegation of disputes to a court as a legitimation.


120 Zuckerberg, supra note 1.

121 Ash et al., supra note 15, at 20.

122 Facebook, supra note 2.
Throughout the development process, one noticeable shift has been the abandonment of the comparison of the FOB to the “Supreme Court.” This is welcome, and this paper argues that the ultimate design should go even further in this direction. Some scholars have suggested that the effectiveness of the FOB “will depend on the answer to one key question: How much will the “Supreme Court of Facebook” be like the Supreme Court of the United States?” Of course, the recourse of commentators and Zuckerberg to the analogy of the United States Supreme Court is not surprising. Facebook is an intrinsically American company, and its staff are acculturated in American culture and legal norms. But focusing on a comparison to the US legal system would be an unduly narrow view. Facebook is a global company. What’s more, since the US Constitution was written nearly 1000 different constitutional systems have been created from which much has been learned. Some of the details in the Draft Charter, described below, show Facebook is open to these lessons and to institutional innovation.

A. Membership

Ultimately, the individuals selected for the FOB will be enormously consequential for the quality of its decisions and the success of the project. But members will be helped or hindered by their institutional context. The Draft Charter announces that the FOB will be comprised of “up to 40 global experts,” selected on the basis of publicly available qualifications as well as geographical, cultural, personal and professional diversity. Deciding the size of the board involves a trade-off: a smaller group of members would create a greater concentration of expertise and authority, while a larger group will be more diverse and may be able to hear more cases. Choosing to have members from varied professional backgrounds also shows this preference for diverse perspectives, instead of selecting members who would have a common set of professional norms and a shared discursive toolbox.

The choice not to prioritize accumulation of expertise and authority will be exacerbated by Board members’ relatively short terms – fixed at three years and automatically renewable once subject to removal for violation of terms of their appointment. Compensation will be set in advance and unchangeable. The fixed tenure and compensation are important – these indicia are “the gold standard of independence” and essential to preserve members’ independence from Facebook. But the short term length does undermine this somewhat. Life tenure (as for US Federal judges and justices) is not necessary, feasible or desirable for Facebook perhaps, but the international norm is for terms of around 9-14 years. The choice of a shorter term not only gives members less time to develop their understanding of their role and the Facebook content moderation ecosystem, but may also mean FOB members are

---

124 Klonick, supra note 38, at 1621.
126 As to professional diversity, the Draft Charter says “The board will be made of experts with experience in content, privacy, free expression, human rights, journalism, civil rights, safety and other relevant disciplines”: Facebook, supra note 2, at 1.
127 Draft Charter at 2.
128 Draft Charter at 4.
129 The Federalist No. 78 (Alexander Hamilton) (“nothing will contribute so much as [permanent tenure of judicial offices] to that independent spirit in the judges which must be essential to the faithful performance of so arduous a duty.”).
130 Mark Tushnet, Judicial Accountability in Comparative Perspective, in ACCOUNTABILITY IN THE CONTEMPORARY CONSTITUTION 57, 57, 61–62 (Nicholas Bamforth & Peter Leyland eds., 2013); David S. Law, How To Rig the Federal Courts, 99 GEO. L.J. 779, 786 (2011) (“It is perhaps typical to think of appointment for life and protection against diminution in salary as the gold standards of judicial independence”); Frank Cross, Judicial Independence, in THE OXFORD HANDBOOK OF LAW AND POLITICS 567 (Gregory A. Caldeira et al. eds., 2008); Vanberg, supra note 89, at 101; Sweet, supra note 86, at 824 (suggesting term of 9-12 years is typical).
unduly concerned with their career prospects post-service.\textsuperscript{131} This perverse incentive to consider their own interests might be further aggravated by the fact that members will only be serving in a part-time capacity.\textsuperscript{132} Indeed, this was promoted as a benefit by Noah Feldman in an early white paper on the FOB, arguing “[r]eputational effects will be important to them. They will be mid-career actors, so that they will have other interests and goals alongside their Facebook judicial position.”\textsuperscript{133} While it is true, as Feldman argues, that this will reduce pressure to break professional norms and engage in attention-seeking behavior, having shorter, part-time terms increases the salience of board members’ personal interests rather than their professional responsibility. These concerns might be somewhat mitigated by the fact that “individual members’ names will not be associated with particular decisions.”\textsuperscript{134} Granting anonymity may protect against pressure to vote in a particular way on any given case, and increase the desire to bolster the legitimacy of the institution as a whole with respect-worthy decisions.

Overall, these choices reflect a theme that dominates the Draft Charter: the importance of diversity. Indeed, the words “diverse” or “diversity” are used six times in two pages.\textsuperscript{135} The reason for this is clear. Facebook has long been criticized for the lack of diversity in its leadership, product design and content moderation teams.\textsuperscript{136} This has affected the company’s ability to anticipate and understand how its platform impacts its diverse global user base. Many of the biggest content moderation controversies have involved the harsher operation of Facebook’s rules on minorities, for example.\textsuperscript{137} Greater diversity enhances the legitimacy of courts’ decisions in the eyes of the public, and especially for those minorities that previously lacked representation.\textsuperscript{138} Kadri and Klonick have argued that “Facebook’s council cannot hope to gain legitimacy if it fails to represent a broad array of viewpoints and experiences.”\textsuperscript{139} Facebook has similarly written that “[t]he success and the ultimate effectiveness of the Oversight Board will depend on its ability to accommodate an inclusive and diverse range of perspectives, across language, culture and

131 Tushnet, supra note 129, at 61–62 (“; if reappointment is precluded, judges may shape their decisions with an eye to their prospects for post-service employment, which may promote indirect accountability to someone, though not necessarily the public. And, for obvious reasons, these effects may differ depending on the age at which judges take their positions on apex courts, and whether they are subject to rules requiring or norms encouraging retirement at a specific age or restrictions on post-service employment.”).
132 Draft Charter at 2.
133 Zoe Mentel Darmé et al., Global Feedback & Input on the Facebook Oversight Board For Content Decisions 143 (Facebook), Jun. 27, 2019.
134 Id. at 4.
135 Draft Charter at 1-2.
137 See, e.g., Mara Schechter, Update: 570k+ people call on Facebook to stop censoring activists, Daily Kos (Mar. 3, 2017), https://www.dailykos.com/story/2017/3/3/1639597/-Update-570k-people-call-on-Facebook-to-stop-censoring-activists (reporting on a letter from civil rights organizations, including the ACLU, criticizing Facebook: “Activists in the Movement for Black Lives have routinely reported the takedown of images discussing racism and during protests, with the justification that it violates Facebook’s Community Standards. At the same time, harassment and threats directed at activists based on their race, religion, and sexual orientation is thriving on Facebook.”); Julia Angwin, supra note 47.
138 See, eg, Nancy Scherer, Diversifying the Federal Bench: Is Universal Legitimacy for the U.S. Justice System Possible?, 105 NW. U. L. REV. 587 (2011); Nancy Scherer & Brett Curry, Does Descriptive Race Representation Enhance Institutional Legitimacy? The Case of the U.S. Courts, 72 THE JOURNAL OF POLITICS 90 (2010); Sweet, supra note 86 (noting also the importance of diversity of expertise).
But this may set an impossible target. The FOB cannot encompass diversity sufficient to represent the views of a user base that is perhaps the most diverse “community” ever assembled. The US Supreme Court has been criticized for failing to reflect the diversity of its single-country jurisdiction. Consider, then, that the FOB’s jurisdiction will cover approximately 191 additional countries. One report noted that “[c]reating a body that has credibility with the extraordinarily wide geographical, cultural, and political range of Facebook users would be a major challenge.” This is an understatement. It is not just that the FOB “cannot hope to represent every single view,” or cannot “include representatives from every country and culture.” It cannot hope to represent even a meaningful fraction of those views. This is especially so given that the 40-member FOB will not sit as a whole in each case, but will instead hear cases in panels. A 40-member body is not even large enough to include a representative from each of Africa’s 54 countries. Facebook acknowledges Africa is a “complex tapestry of cultures,” but then proceeds to flatten this diversity by saying “Africa will be represented on the board alongside other regions.” There is no reason to believe, for example, that any representative from Africa will have the expertise and legitimacy necessary to authoritatively pronounce on issues as diverse as Nigerian fake news to Libyan “keyboard warriors” who carry out online partisan battles.

Rather than pretending that the FOB itself can be sufficiently diverse to be representative of Facebook’s users, the limits of this capacity should be factored into the FOB’s design. There is no single body that could represent a community the size and scale of Facebook. The FOB will instead need to be empowered to reach out to different experts and communities who have relevant perspectives to offer in any individual case. Facebook is not wrong that the legitimacy and therefore success of the FOB will depend in large part on its capacity to bring greater diversity to Facebook’s content moderation system; but it will need to find ways to do this other than through FOB membership alone.

### B. Power of Review

The Draft Charter states that the “primary function of the board is to review specific decisions we make when enforcing our Community Standards.” As noted above, this may be the primary activity of the FOB, but it cannot be its primary function. Given the sheer volume of content moderation decisions that Facebook makes every day, the FOB’s impact would be miniscule if its function were purely error correction in individual cases. Other aspects of the Draft Charter suggest Facebook appreciates this broader purpose.

---

140 Archibong, supra note 112.


142 Facebook is blocked in China, North Korea and Iran.

143 Ash et al., supra note 15, at 20.

144 Kadri & Klonick, supra note 139, at 42.

145 Archibong, supra note 112.

146 Draft Charter at 3.

147 Archibong, supra note 112.


150 Draft Charter at 3.
The FOB’s decisions will be made public and include an explanation.151 If the decision of the relevant panel is not unanimous, a member can also release a dissenting opinion.152 This shows that the purpose of the FOB is not only to decide outcomes and correct errors, but also to give reasons: to expose the reasoning and tensions involved in content moderation decisions.

To guide this process, Facebook will publish a final Charter that includes a set of values which the FOB agrees to uphold.153 This aims to make the basis for the FOB’s decisions more transparent, thereby ensuring “the public legitimacy of the board will grow.”154 FOB panels should “ensure consistency with other issued opinions”155 so that its decisions create a coherent body of “platform law.”156 This emphasis on consistency and coherence with prior decisions invokes notions of stare decisis, the legal doctrine of precedent. Given this reliance on a fundamental concept of legal training, it is again notable that not all members of the FOB will be lawyers.

Given that reason-giving is central to the FOB’s role, its institutional design should be centered around facilitating the production of quality decisions. Again, this involves trade-offs, the most obvious being speed of decision-making in order to ensure that by the time the case is decided there is still the possibility of a substantive remedy. If a take-down decision comes too late, the damage of a viral post may already be done. The importance of speed versus careful consideration may differ in each case, but the Draft Charter indicates there will be a hard deadline of two weeks for all FOB decisions.157 The difficulty of this rigid timeline is indicated by the wide range of activities that the FOB might undertake during this time: it will be empowered to “call upon experts to ensure it has all supplementary linguistic, cultural and sociopolitical expertise necessary to make a decision,”158 and other stakeholders will also be able to submit arguments and material to the panel.159 This additional, inquisitorial-style evidence gathering powers of the FOB are an essential departure from the US Supreme Court model. The factual record on which Facebook’s content moderators base their decisions is incredibly slim—a screenshot of the post in question.160 As Kadri and Klonick have noted, content moderation decisions are highly context dependent:161

“[T]he difference between a racist slur and a rap lyric, for example, might turn on the speaker’s identity, her motivations, her audience. These challenges become even more complex in a global context in which moderators must account for different languages and slang; for different historical, cultural and political divides; and for different power structures — all of which might color the social meaning of the speech.”

Zuckerberg has also acknowledged that the “linguistic nuances” involved in identifying hate speech and bullying pose a special challenge for both reviewers and AI.162 The FOB’s role will therefore be especially important in these cases.

In difficult cases, then, the extent of the material necessary for the FOB to make a well-informed decision

---

151 Draft Charter at 5.
152 Draft Charter at 5.
153 Draft Charter at 5.
154 Draft Charter at 5.
155 Draft Charter at 5.
156 Kaye, supra note 13, at 3.
157 Draft Charter at 5. Although it is not clear from the Draft Charter when this clock will start running.
158 Draft Charter at 4.
159 Draft Charter at 4.
160 Klonick & Kadri, supra note 123.
161 Id.
162 Zuckerberg, supra note 1.
could be vast, or the logistics of finding and hearing from necessary experts might be challenging. But because of the limitations on the FOB’s ability to adequately represent the full diversity of Facebook’s users, as noted above, this expertise-gathering should be given ample space to occur. In this context, the two-week deadline seems somewhat arbitrary: a functional board would itself determine the correct balance in each case. It is unclear why a strict standard of two weeks strikes a good balance—a fortnight is an age in terms of the internet zeitgeist (justice delayed is virality denied) but perhaps not long enough for a multi-member board to gather and consider all the materials it needs.

Another notable departure from the US model is the FOB being given a kind of “abstract” jurisdiction. Facebook has indicated that beyond deciding individual cases, the company “may request policy guidance from the board.”163 The US Supreme Court’s jurisdiction hinges on the finding of a specific “case or controversy,”164 but an abstract jurisdiction for the FOB will empower it to highlight problems in Facebook’s rules in advance of individuals bringing appeals about them. This capacity could be further enhanced by the creation of a concrete review jurisdiction,165 where moderators might escalate matters to review in hard cases where they are uncertain about how the Standards should be applied, in advance of a user-initiated “appeal.” This would have the advantage of potentially headng off controversy created by moderators being forced to make a decision in hard cases without the benefit of full consideration. The power to assess the appropriateness of rules and not merely the application of them in a particular case is an important one, discussed further below, and should be available beyond Facebook’s discretion to refer such cases.

Many questions remain about how, in practice, the FOB will exercise these powers of review. Not least among them are “standing” rules about who is qualified to bring an appeal (just the user who posted the content, for example, or third parties?), the form of argumentation (will there be a Facebook bar who argues before the FOB in person?), how cases will be selected from the large volume of contested decisions, how to support “litigants” to ensure that access to the FOB is not limited to the well-resourced, who will bear the burden of proof to establish error, what kind of evidence the FOB will consider, and innumerable other questions. These decisions will impact the FOB’s legitimacy, and many will likely be worked out over time. I am concerned here with institutional design of the FOB at a higher level of generality, and so I set these questions aside for now. But once the FOB is running, these questions will become more urgent and salient in how it is perceived.

C. Subject-Matter Jurisdiction

For the FOB to meaningfully contribute to Facebook’s content moderation ecosystem, its “subject-matter jurisdiction” — that is, the topics on which it is empowered to hear cases — should reflect the wide ambit of ways Facebook decides what appears on its platform.

Two things are worth noting about the subject-matter jurisdiction of the FOB. First, the FOB “will not decide cases where reversing Facebook’s decision would violate the law.”166 This reflects Facebook’s general position that it will respect local laws in the countries it operates.167 Facebook follows local laws because they “are often the result of public input, even just the indirect influence of a democratic election. The laws are therefore likely to reflect, at least in democracies, the social values of the local

163 Draft Charter at 3.
164 United States Constitution, Art. III, s 2, cl 1. This is a common feature in the ideal type of a constitutional court: Sweet, supra note 86, at 818.
165 SWEET, supra note 8, at 45.
166 Draft Charter at 3.
167 Bickert, supra note 29, at 258.
population."168 This restriction on the FOB’s jurisdiction is not surprising: indeed, where local law conflicts with Facebook’s rules, the FOB would have limited power to mandate that Facebook disobey legal requirements. But a consequence of this is that governments have de facto control over the FOB’s remit. A government can prevent the FOB from giving an opinion on a matter if it makes a particular type of content illegal. As Feldman acknowledges, “It would be plausible to eliminate this provision from the hence remit. A government can prevent the FOB from giving a particular opinion on a matter if it makes a particular type of content illegal. As Feldman acknowledges, “It would be plausible to eliminate this provision from the proposal altogether. The idea would be to discourage states from enacting laws that limit expression and hence gaining control over content on Facebook.”169 This provision also undermines Facebook’s professed desire to have a universal set of Community Standards. In practice, and particularly in countries where governments seek to exercise tight control over public discourse, compliance with local law may be a severe limitation. National and international laws about free speech and online content are proliferating and diverse.170 Even within a single jurisdiction, different courts can come to different conclusions on the meaning of relevant laws and whether certain content is illegal.171

A key anterior question in these cases will be who decides whether a take-down decision was taken due to legal requirements or Facebook’s own Community Standards. If Facebook itself decides this question, this too undermines the FOB’s oversight capacity by giving Facebook a way of deciding that certain cases should not reach the FOB. Facebook could determine, without the possibility of review, that the FOB cannot hear a particular case. This may be exacerbated by the fact that when legal liability is not clear, platforms tend to err on the side of caution.172

The second notable thing about the subject-matter jurisdiction of the FOB is its overall narrowness. Apart from possible abstract review referrals by Facebook, the FOB is limited to reviewing individual applications of Facebook’s Community Standards. The biggest controversies around content moderation for Facebook in the last few years have concerned its decisions to take down, or not take down, putative hate speech,173 foreign interference in domestic elections,174 and other forms of misinformation and disinformation (including so-called “fake news,”175 “deepfakes” and “cheapfakes”176). But these are not the only important types of content moderation decisions Facebook makes. Allowing tech companies to frame “content decisions” or “content moderation” in such a limited way would get in the way of creating

---

168 Id.
169 Darmé et al., supra note 133, at 142. (emphasis added)
meaningful oversight and instead result in a type of transparency theatre.

As examples, there are at least two other key areas of “content decisions” that the FOB should have jurisdiction to adjudicate disputes under, both of which would be a more significant abdication of authority by the company. The first is algorithmic ranking decisions, and in particular the decision to down-rank certain content and decrease its circulation. The second is the application of its advertising policies, especially those around compliance with the additional new requirements Facebook has put in place concerning political ads.

Both of these decisions go much more directly to the core of Facebook’s business model than individual content moderation decisions about whether user posts comply with the Community Standards. The Facebook News Feed algorithm is Facebook’s “secret sauce,”177 that drives user engagement. Facebook is notoriously protective of its algorithm,178 and users,179 academics,180 civil society,181 and lawmakers182 have long been calling for greater transparency around how Facebook determines what gets shown in the News Feed. And advertising is the backbone of Facebook’s business.183 Facebook might be concerned that giving the FOB power to review decisions in these domains risks too much intervention in the platform’s core product design and revenue stream. But denying this jurisdiction undermines any claim that the FOB is a meaningful and bona fide attempt to give greater rigor and transparency to Facebook’s content moderation ecosystem as a whole. Indeed, confining the scope of jurisdiction of courts is a common technique used by authoritarian regimes to undermine the check on power that an independent judiciary might otherwise provide.184 As Martin Shapiro has observed, “[a] relatively independent judiciary may be preserved but simply excluded from domains significant to the authoritarian regime.”185

---

177 THE FACEBOOK DILEMMA, Frontline, Part I at 12:22 (Antonio Garcia Martinez, former product manager at Facebook).
184 Moustafa, supra note 34; Law, supra note 130, at 812 n150 (describing how Taiwan’s legislative body changed the quorum requirements for the issuance of constitutional interpretations in order to substantially diminish such rulings); CARLO GUARNIERI & PATRIZIA PEDERZOLI, THE POWER OF JUDGES: A COMPARATIVE STUDY OF COURTS AND DEMOCRACY 78–79 (Cheryl Thomas trans., Oxford Socio-Legal Studies, 2002) (“the kind of disputes a judge can be asked to settle … provides the starting point for assessing the scope of courts. … Authoritarian regimes often adopt [a] fragmentation strategy to control the administration of justice.”); Landau & Dixon, supra note 77, at 30.
185 Martin Shapiro, Courts in Authoritarian Regimes, in RULE BY LAW: THE POLITICS OF COURTS IN
As I show next, both ranking and advertising decisions are core content moderation matters, and a body that is intended to provide “oversight” to content decisions needs to be empowered to review them.

A Degree of Algorithmic Transparency

Often, content moderation conversations revolve around a “take down / leave up” dichotomy. But platforms have far greater capacity to control the content on their sites than this paradigm suggests. Facebook is increasingly relying not on the blunter content moderation tools of removing posts or pages, but on the subtler tools of limiting their reach and exposure. For “borderline” content in each of its harmful categories of content, Facebook works to “distribute that content less” to reduce the incentive to post such content.186 Zuckerberg argues that “no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average.”187 But Facebook’s decision to “downrank” a piece of content (or distribute it less) in users’ News Feeds is currently much less transparent than a decision to take down a piece of content. Users are typically notified when a post is removed entirely, but may be entirely unaware when their post is left up but just not shown to other users. From Facebook’s perspective, this is a key virtue of this approach. As the company notes in a recent patent for this practice, when it is used “the commenting user is not made aware that his or her comment was blocked, thereby providing fewer incentives to the commenting user to spam the page or attempt to circumvent the social networking system filters.”188

Opacity in Facebook’s ranking decisions leaves users guessing at what is happening, and speculating if this is the cause of any drop in engagement with their posts. Some (including President Trump) have suggested that social media platforms “shadow ban” posts simply because they do not like them.190 Crucially, if the FOB cannot review down-ranking decisions not only will this speculation persist, but also Facebook can still effectively control the extent to which the FOB can provide oversight: if Facebook does not want a takedown to be subject of public attention and the FOB’s review, Facebook could simply downrank that content effectively out of circulation instead of removing it.

Facebook’s content moderation includes decisions about how to distribute content Facebook does not take down. Therefore, to give legitimacy to this wider ecosystem the FOB needs jurisdiction over these decisions. Of course, in a sense, every piece of content on Facebook is subject to a degree of algorithmic content moderation. Facebook has famously jealously guarded the workings of its News Feed algorithm, to prevent user gaming and for trade secrets reasons. There may be a legitimate argument that the FOB cannot or should not have complete oversight of all aspects of the algorithm; but at least to the extent that algorithmic changes are used explicitly as a tool of content moderation on the basis of lines drawn in the community standards, these should be subject to FOB review.

Advertising Policies

Facebook’s has detailed policies on advertising and all ads purchased undergo pre-publication review to ensure they meet Facebook’s rules.191 Facebook has made significant changes to its political advertising

AUTHORITARIAN REGIMES 326, 331 (Tamir Moustafa & Tom Ginsburg eds., 2008).

186 Zuckerberg, supra note 1.
187 Id.
191 Advertising policies, Facebook, https://www.facebook.com/policies/ads#.
policies in the wake of controversy surrounding the 2016 US election. Advertisers that run political ads are now required to go through an authorization process and run disclaimers about who paid for the ads. Facebook’s definition of a political ad includes an ad that relates to any of 20 “issues of national importance,” a list that includes, for example, the indeterminate category of “values.” Facebook’s application of these rules has caused problems. Ads from news sites and non-profits were removed because they were flagged as “political” but had not received the necessary special authorization, while ads purchased in the names of sitting Senators were waved through. The individual and societal consequences of inconsistent and opaque enforcement of political advertising policies is potentially significant. Yet, as Kreiss and McGregor found, Facebook “exercise[s] considerable discretion over paid political media. While these decisions are clearly difficult, they are made without much in the way of transparency, consistency, or accountability.” The lack of explanation for how these policies are applied is particularly concerning because without it “Facebook would have the authority to determine the messages that campaigns pay for the public to see, without the [public] challenges that keep these firms learning and cause their policies to evolve.” Without transparency, it also opens Facebook up to the charge that its decisions are politically motivated.

Inconsistent and unexplained decision-making is exactly the problem the FOB is being created to solve. In the context of political content, this is especially important. The mere fact that content is paid and not “organic” does not alleviate the need for oversight. Giving the FOB jurisdiction in these cases would put literal money where Facebook’s mouth is about trusting the expertise of the FOB to interpret and apply its rules in a legitimate manner.

In the report summarizing the results of its global consultation on the FOB, Facebook said that it had been “clear” that the Board was only going to review individual decisions under the Community Standards and that while News Feed ranking and political advertising were “important issues,” they were beyond the remit of the Board. But these are not just important issues: they are at the core of Facebook’s content moderation ecosystem. Excluding them from the FOB’s jurisdiction undermines Facebook’s broader commitment to transparency and accountability in content moderation that the FOB is intended to facilitate.

192 Id. at ("5. Restricted Content, 11.a. Ads related to politics or issues of national importance").
198 Id. at 11–12.
199 Darmé et al., supra note 133.
IV.  THE LEGITIMACY CONUNDRUM

The fundamental purpose of the FOB is to bring greater legitimacy to Facebook’s content moderation ecosystem. While enhancing the legitimacy of the FOB itself is instrumental to this goal, it is not the ultimate aim: the aim is to legitimize Facebook’s exercise of power by subjecting it to an independent check. The focus of this essay is therefore this more general sociological legitimacy – that is, the extent to which the public regards Facebook’s decisions generally as justified, appropriate, or otherwise deserving of support beyond the fear of sanctions or mere hope for personal reward.200

In a crucial sense, this legitimacy can only be established over time. For the FOB itself, as the Draft Charter notes, “[t]he public legitimacy of the board will grow from the transparent, independent decisions that the board makes.”201 But gaining legitimacy will be difficult. The FOB is being established in response to widespread public dissatisfaction with Facebook content moderation and so the enterprise begins on the back foot. And while the early use of the moniker of a “Supreme Court” may have been (intentionally or unintentionally) intended to invoke the aura of legitimacy which the United States Supreme Court enjoys, the FOB will not have a reservoir of legitimacy accumulated over a long history to draw down upon.202

But there are two more fundamental challenges in the quest for legitimacy: first, the difficulty for Facebook making a credible commitment to being bound by the FOB when, ultimately, Facebook itself retains final authority; and second, the lack of agreed prior norms or authorities for content moderation which the FOB can base its decisions on and leverage into accepted pronouncements. The rest of this part addresses these in turn.

A. The Limits of the FOB’s Legitimizing Power

There is a real limit on the extent of the legitimation that the FOB can bring to Facebook content moderation. Above, I discussed the ways in which authoritarian regimes use courts to give them a veneer of legitimacy. But as Mark Tushnet has pointed out about this literature, “[t]he general difficulty with these accounts is straightforward: rulers might want to make credible commitments, but they cannot do so, precisely because they can alter the constitution whenever they want—and the target audiences know that the rulers can do so.”203 In Facebook’s case, the conundrum is the same: Zuckerberg wants to assure stakeholders that the FOB will be independent, make decisions “in the best interests of our community and not for commercial reasons,”204 and that Facebook will consider itself bound by the FOB’s decisions. But it is difficult to make this commitment to tie itself to the mast of the FOB credible. Ultimately, Facebook does not have to obey the FOB’s decisions or could disband the institution altogether without cause. This is the flip side of judicial review being a relatively low-cost way of enhancing legitimacy without opening up the broader policy-making process – you get what you pay for.

A practical issue may also arise for Facebook of how to resolve its fiduciary duties to its shareholders should the FOB issue a decision that it genuinely regarded as against its commercial interests.205 The

---

201 Draft Charter at 5.
203 Tushnet, supra note 88, at 422–23.
204 Zuckerberg, supra note 1.
simple response is that the FOB can be considered in the long-term commercial interests of the company because it increases user satisfaction, protects company reputation and potentially staves off more severe regulation. But to the extent that this is true, the FOB does not actually constitute a “check” on Facebook’s power because its actions remain in Facebook’s best interests. As Tushnet explains, where it serves the interest of the authoritarian for there to be an external check on its powers, “courts do no work because the regime’s immediate self-interest will lead it to refrain from actions that reduce the returns it anticipates to gain during the period when the preferences are stable. And, if the regime’s preferences change, the mechanisms also do no work because the regime is free to change them to accommodate its new preferences.”\textsuperscript{206} Similarly, should Facebook decide that the cost of public legitimacy becomes too high, it can simply abandon its experiment.\textsuperscript{207}

The notion that the FOB is merely instrumental to Facebook and does not constitute a meaningful check is further suggested by a number of design choices. The highly circumscribed and subjective nature of the FOB’s subject-matter jurisdiction, discussed above,\textsuperscript{208} is the most significant. But there is also the fact that Facebook has said that the company “\textit{can} incorporate the board's decisions in the policy development process,”\textsuperscript{209} suggesting that this possibility is discretionary. There is a fundamental tension for Facebook in creating a meaningfully independent body while also keeping its discretion within acceptable bounds. As Zuckerberg asked, “How do we ensure their independence from Facebook, but also their commitment to the principles they \textit{must} uphold?”\textsuperscript{210} Facebook’s answer to this has been to require FOB members to agree to uphold a set of values specified by Facebook.\textsuperscript{211}

\textbf{B. The Limits of a “Constitution’s” Legitimizing Power}

Requiring FOB decisions to be based on a set of underlying values has been likened to the adoption of a “Constitution” which will guide the FOB’s interpretation of the Community Standards.\textsuperscript{212} These values will play an important role in the FOB’s work, but there are crucial ways in which they are unlike a Constitution. A values statement can perform the same role as Constitutions do in expressing Facebook’s fundamental vision for its platform which the FOB can use to resolve ambiguity in the Community Standards in a way that reflects these commitments. But because the values will be Facebook’s and not the user community’s, they cannot perform the same role of legitimation that Constitutions do because there is no sense in which the values express a delegation of authority from users or a set of widely-agreed norms.

The essential role that the statement of values will play is guide the FOB in difficult cases where there is no clear answer. It is precisely because there will be instances where competing understandings of the Community Standards are possible that the FOB is necessary. If the Community Standards provided a clear answer in every case, there would be no need to seek review of content moderation decisions. In cases where there are multiple possible interpretations of these rules, what should the FOB use to decide between them? Without a set of underlying values, decisions become relatively unconstrained. This is suboptimal from Facebook’s perspective, because it creates a large sphere of discretion for FOB

\textsuperscript{206} Tushnet, \textit{supra} note 88, at 425; see also Khan & Pozen, \textit{supra} note 205, at 10.
\textsuperscript{207} For discussion of the fact that the FOB represents a “bet” that public legitimacy is a worthwhile commercial investment that could be abandoned, see Evelyn Douek, \textit{YouTube’s Bad Week and the Limitations of Laboratories of Online Governance}, Lawfare (Jun. 11, 2019), https://www.lawfareblog.com/youtubes-bad-week-and-limitations-laboratories-online-governance.
\textsuperscript{208} See \textit{XX}.
\textsuperscript{209} Draft Charter at 1 (emphasis added).
\textsuperscript{210} Zuckerberg, \textit{supra} note 1 (emphasis added).
\textsuperscript{211} Draft Charter at 5.
\textsuperscript{212} Kadri & Klonick, \textit{supra} note 139, at 43.
members. But it is also suboptimal for users and for the legitimacy of the FOB itself: without an underlying set of commitments, decisions can become unpredictable or seen as arbitrary.

In its Draft Charter, Facebook said it would adopt a set of values that “would encompass concepts like voice, safety, equity, dignity, equality and privacy.”213 Importantly, a list of vague “values” does not help achieve the aim of guiding the FOB’s interpretative practice. A list that includes everything prioritizes nothing. It is hard to disagree with the importance of each of “voice, safety, equity, dignity, equality and privacy,” but the very nature of hard freedom of speech cases is that they involve trade-offs between these values. The difference between various countries’ hate speech jurisprudence, for example, is largely due to a different weighting of these values. As Rosenfeld explains:

If free speech in the United States is shaped above all by individualism and libertarianism, collective concerns and other values such as honor and dignity lie at the heart of the conceptions of free speech that originate in international covenants or in the constitutional jurisprudence of other Western democracies. Thus, for example, Canadian constitutional jurisprudence is more concerned with multiculturalism and group-regarding equality. For its part, the German Constitution sets the inviolability of human dignity as its paramount value … These differences have had a profound impact on the treatment of hate speech.214

Albie Sachs has similarly observed:215

To Americans, the firstness of the First Amendment is axiomatic. It is seen as a source of enlightenment, as being the most constitutive and defining element of the whole constitutional order. The legal cultures of Germany and South Africa, however, have a profoundly different foundational element. It is not free speech, but human dignity. What is axiomatic to an American lawyer could be problematic to us. What is axiomatic to us could be problematic to an American.

Just as AI cannot perform content moderation without coders making value judgments in writing the algorithms,216 so too Facebook cannot escape the need to make a choice about the kind of platform it wants to be. The decision of whether to allow conspiracy-monger Alex Jones to continue to stream and share violent and sometimes racist conspiracy theories is an example where such a choice is necessary.217 On one side is the value of voice and liberty to speak, but on the other hand there is the safety and dignity of people affected by his hateful rhetoric or conspiracy-mongering. Another example is the decision to prohibit content promoting eating disorders, which reflects a prioritization of safety over unconstrained voice.218 These choices are controversial – this is part of the reason the FOB is being set up. But an uncontroversial statement of values that is universally acceptable would be stated at such a level of generality to be of limited utility in constraining decision-making.

213 Draft Charter at 5.
216 See Part II(A) above.

27
Zuckerberg has in fact long articulated underlying values he sees as inherent in his platform. In the opening words of his Blueprint announcing the FOB’s creation he says “[m]any of us got into technology because we believe it can be a democratizing force for putting power in people's hands. I've always cared about this and that's why the first words of our mission have always been "give people the power".”\textsuperscript{219} This kind of rhetoric attracts backlash, but a transparent statement of what the platform is intended to be is important in explaining choices made in content moderation. And by explicitly choosing priorities in advance, this reduces the chances that the FOB’s decision-making becomes so out of step with the company’s vision for its platform that it decides to simply ignore or overrule its decisions.

But although a Charter that makes difficult choices will be necessary for the FOB to function effectively, it cannot cure Facebook’s legitimacy deficits for several reasons. First, as discussed, there is no mechanism to double-entrench any “constitution” such that Facebook could not amend it at will, making it difficult to place too much faith in its power. Second, the constitution itself will not have sociological legitimacy and so users have no special reason to accept decisions made ostensibly in accordance with it. A list of human rights buzzwords released by Facebook or chosen by Zuckerberg will not have the same resonance as national constitutions that are typically adopted through democratic processes. Although Zuckerberg seems to have been reading some Rawls,\textsuperscript{220} he is more founder than founding father and has little mandate to decide the values of the entire community of Facebook users. A choice by him to prioritize “voice” over “dignity,” for example, might resonate with some users but cannot be said to be reflective of the will of those users in any real sense nor will those that disagree have much reason to feel themselves reasonably beholden to Zuckerberg’s choice.

Facebook could, and many have argued should,\textsuperscript{221} adopt other widely endorsed norms such as international human rights law as the basis of its decisions. Facebook itself has said it looks to documents like the International Covenant on Civil and Political Rights for guidance on where to draw the lines on freedom of expression.\textsuperscript{222} There are some potential limitations with adopting these norms wholesale,\textsuperscript{223} but they could provide a starting point on which to develop a more mature form of content moderation. International human rights law has the advantage of having a significant reservoir of legitimacy in certain communities which Facebook and the FOB could draw down on. Many scholars are doing important work on this topic, but I set it aside here for the simple reason that at this stage Facebook is showing no indication that it intends for international human rights law to be the fundamental basis for its content moderation.

Therefore, a fundamental set of values is a necessary but not sufficient part of creating legitimacy for Facebook’s content moderation and the FOB’s decisions.

\textsuperscript{219} Zuckerberg, supra note 1.

\textsuperscript{220} Highlights & transcript from Zuckerberg’s 20K-word ethics talk | TechCrunch, https://techcrunch.com/2019/02/20/zuckerberg-harvard-zittrain/.


\textsuperscript{223} This is discussed in Douek, U.N. Special Rapporteur’s Latest Report on Online Content Regulation Calls for “Human Rights by Default,” supra note 221 I will explore these issues further in later work.
V. MAXIMIZING THE FOB’S POTENTIAL

The FOB can still bring significant benefits to Facebook’s content moderation despite these limits on the legitimacy it can create. Maximizing these benefits requires acknowledging these limitations and designing around them. This includes accepting two facts inherent in the problem of content moderation on private platforms: Facebook will always retain the power to overrule the FOB’s decisions, and in many of the most difficult cases there will be no “right” answer. Institutional design that factors these facts in will be stronger in the long run. It should focus on the two key benefits that the FOB can bring. First, a practical benefit: a judicial-style check, even if reversible by Facebook, can improve Facebook’s policy formation processes. Second, a sociological benefit: bringing greater public acceptance of Facebook’s rules through the performance of public reasoning.224

A. Weakness as Strength

The FOB can improve Facebook’s policies without being the ultimate authority on every aspect of the platform’s rules. The idea that the FOB’s review of Facebook’s decisions should be final and irreversible, not only in the particular case but in all similar cases going forward, is a fundamentally American perspective. Indeed, this is known as strong-form judicial review, and the United States Supreme Court is considered the “archetype or paradigm” of this practice.225 Tushnet, the leading writer on the strong-form/weak-form distinction, describes the strong-form model in the following way:226

In systems with strong-form judicial review, of which the United States is usually taken to be representative, the constitutional court has the power to invalidate primary legislation, and legislative responses to such invalidations are made quite difficult.

By contrast, the central characteristic of weak-form review is that “courts assess legislation against constitutional norms, but do not have the final word on whether statutes comply with those norms” and judicial interpretation can be displaced by ordinary legislative processes.227 Commentators have suggested that this kind of override would undermine the FOB’s independence.228 Kadri has argued that “[i]f Facebook is free to unilaterally overrule appellate decisions it doesn’t like, the talk of a new era of radical transparency and accountability will be overblown. . . amending its legal code willy-nilly would undermine the entire project.”229 But the comparative literature on “weak-form” or “dialogic” review suggests this may be too simplistic. There are many benefits of a weak-form model of review that may make it more appropriate for the dynamic and complex environment that is Facebook.

224 This Part expands on some of the arguments made in Douek, Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Governance, supra note 119.
226 Mark Tushnet, Judicial Activism or Restraint in a Section 33 World Review Article, 53 U. TORONTO L.J. 89, 89 (2003).
229 Kadri, supra note 228.
1. The Benefits of a Judicial-Style Check on Policy-Making

As Dixon argues, weak-form review is well-suited to counter blockages in the “legislative process” (here, the formulation of Community Standards). Dixon describes two forms of blockages: blind spots and inertia. Both are present in Facebook’s policy formation.

Blind spots arise because initial policies are often written in time-pressured conditions, and cannot fully anticipate all the possible circumstances in which a policy will need to be applied. This well describes the process for the formulation of Community Standards at Facebook. Initial rules were written haphazardly as international expansion meant the platform needed to accommodate a rapidly growing and changing user base, and revisions have often been prompted by particular high-profile controversies. As Klonick has observed, “internal policies and the rules that reflect them are constantly being updated … because Facebook is attempting, in large part, to rapidly reflect the norms and expectations of its users.” This description of the reactive nature of Facebook’s policy making parallels the description given by Calabresi, who notes that blind spots can arise because “Legislatures often act hastily or thoughtlessly with respect to fundamental rights because of panic or crises or because, more often, they are simply pressed for time.” Judicial-style checks help overcome blind spots because they consider the application of rules retrospectively and in the context of particular cases.

Inertia arises where policy-makers do not have the time to devote to changing rules that are already in effect or because the status quo will often prevail in hard cases when there are disagreements about what should be done. This can be seen in the history of content moderation on Facebook too. For example, the status quo of leaving up Alex Jones’ posts prevailed until the company was forced to reconsider as a result of other platforms making a decision to remove him. The FOB can create a more formalized process for prompting Facebook to address blockages. The process of review can disrupt the status quo, draw public and Facebook’s attention to blind spots or inertia, therefore providing an opportunity for Facebook to correct any error.

This form of check also capitalizes on various actors’ specializations. FOB members will have expertise in “content, privacy, free expression, human rights, journalism, civil rights, safety and other relevant disciplines.” But Facebook staff will have better understanding of the overall dynamics of

230 Dixon, supra note 25.
231 Id. at 2208–9.
232 Klonick, supra note 38, at 1630–35, 1648–58; Gillespie, supra note 27, at 66.
233 Klonick, supra note 38, at 1649.
237 Dixon, supra note 25, at 2209–12.
239 Dixon, supra note 25, at 2216–19.
241 Draft Charter at 1.
speech on its platform and the surrounding architecture. A FOB empowered with a weak-form judicial review mandate can create an ecosystem that puts these two kinds of expertise in “dialogue.”

The FOB does not need to have ultimate authority over Facebook’s rules to bring these benefits. First, even without such authority, Facebook overruling the FOB is likely to be rare. In practice, there would be significant reputational costs for Facebook in disregarding a decision of the FOB. Because Facebook has made a significant investment in publicizing the FOB and extolling its benefits, there would likely be significant attention given to any significant undermining of its authority, and so Facebook is unlikely to take the decision lightly. If done often, it would give rise to the impression that the FOB is mere window-dressing and deprive it of all legitimacy. Furthermore, it is important to remember that a key benefit of the FOB from Facebook’s perspective is not the substantive rulings the FOB issues but the ability this gives Facebook to distance itself from controversial decisions. This incentive remains even where Facebook could overturn any decisions. The similar desire to blame courts for unpopular decisions means that in systems with weak-form review (such as the UK) political actors often use rhetoric that suggests they consider themselves bound by court decisions in order to pass the political buck.242 Therefore the practical strength of “weak-form” review is often much stronger than it appears in theory. As Feldman wrote in his original argument for a FOB, “As a corporation, [Facebook] has the right and capacity to change its policy at any time or even ignore it. The public understands this, as it should. … [But] [i]f Facebook were to violate its commitment, it would be subject to public censure and criticism.”243

Additionally, in cases where Facebook does decide to overrule the FOB, it is likely to issue an explanation given the public attention that would surround the move. This reason-giving is also valuable, and not only as an improvement over the current (lack of) transparency that attends many changes in Facebook’s policies. In explaining why Facebook took a different path to the one recommended by the FOB, this could improve the quality of debate around Facebook’s rules more generally. Currently, contentious content decisions are made in ways that do not facilitate productive discussion, such as executives deciding matters in the early hours of the morning based on media reports of what other tech companies are doing.244 Introducing a judicial style body into a system can elevate the mode of decision-making and discourse, “engender[ing] new modes of legislative discourse and practice.”245 Mattias Kumm calls this “Socratic Consetation."246 By putting decision-makers in dialogue with a judicial-style check, they begin to think about how their decisions can be justified in terms that are more publicly acceptable.

2. The Advantages of Weak-Form Review

So far I have addressed the reasons why weak-form review will in most cases be sufficient to bring about the benefits of an oversight body. But there are also reasons why this model might in fact be preferable to strong-form review. Here, I address five: two theoretical, and three pragmatic. Although less august, the pragmatic considerations may ultimately be more important.

First, weak-form review is more appropriate in cases that involve competing rights claims. Some proponents for judicial review acknowledge that strong-form review might be less appropriate when the judicial body will frequently have to decide zero-sum controversies involving the collision of two

242 Kavanagh, supra note 225, at 1022, 1027.
243 Darmé et al., supra note 133, at 140.
244 Roose, supra note 238.
fundamental rights.247 This is because such cases involve difficult trade-offs that are better left to a branch of government that is more democratically accountable or responsive. The speech disputes that the FOB will be charged with deciding, such as whether hate speech or sexually explicit materials should be taken down, are “common and readily expressible as ‘zero sum’ situations.”248 These are archetypal clashes between liberty rights (to speak) and dignity or equality rights. While Facebook itself is not democratically accountable like a legislature, it may still be more responsive to public pressure than a purposefully insulated FOB. Importantly, Facebook is also able to take proactive action in response to such pressure in the event that rules or FOB rulings become out of step with community values, rather than the FOB which will only make rulings in response to cases brought to it.

Second, because these cases involve such difficult trade-offs, where ultimately one person’s right or interest must give way to another’s, it is important to keep in mind the fragile legitimacy of the FOB itself. In such cases, the lack of a reservoir of goodwill on which to draw on to bolster decisions that are controversial and will upset large segments of the user base will be particularly noticeable. This is especially so given, as discussed above, the FOB’s inability to be truly or comprehensively representative. This means that it will necessarily be making decisions to trade-off the rights or interests of people who will not feel they are represented in the decision-making process. While this can be somewhat mitigated by creating processes through which these communities are given a voice in proceedings, it will never be wholly resolvable. As such, it is better to be humble about the capacities of the FOB to be the final arbiter of norms for all of Facebook’s diverse community and allow for feedback and development.

Third, and this is the critical pragmatic point, if the FOB’s decisions are not seen as final and irrevocable, this may make Facebook more willing to give the FOB the broader jurisdiction that I have argued above is necessary for it to be meaningful oversight of Facebook’s content moderation as a whole. A narrowly confined and manipulable jurisdiction will significantly impair the FOB’s value. It is perhaps the single biggest threat to the FOB’s legitimacy because it creates the impression that Facebook is only willing to renounce power in areas that do not “really matter” to it. However, it is also understandable that Facebook may not want to completely renounce power over this central aspect of its platform. Facebook’s content moderation is in many ways its distinctive value offering.249 Facebook’s right to determine the content that appears on its platform may also be expressed this in terms of the “free speech” rights of platforms themselves.250 Weak-form review may be a concession to this consideration, but also to encourage the benefits of dialogue between Facebook, the FOB and stakeholders over a broader range of areas. This includes the FOB’s ability to influence policy and not just rule on specific take-downs. This kind of abstract review jurisdiction is common in countries with specialized constitutional courts.251 This is because these courts were created, “explicitly and as a constitutional priority, to protect rights” and limiting the avenues to access judicial review would produce gaps in rights protection.252 And even in the US there is a partial exception to the concrete case or controversy requirement for First Amendment challenges brought against laws for being vague or overbroad, given the potential chilling effect of such legislation.253 This is a recognition of not making the ability to bring challenges to speech rules too narrow or difficult. Therefore, a generally broad jurisdiction, including the ability to bring abstract review cases, is critical. If Facebook is more likely to create this jurisdiction and make use of it more often if it

247 Fallon, supra note 236, at 1731; Mark Tushnet, How Different Are Waldron’s and Fallon’s Core Cases For and Against Judicial Review?, 30 OXF. J. LEG. STUD. 49, 55 (2010).
248 Tushnet, supra note 247, at 55.
249 GILLESPIE, supra note 27, at 13.
251 Jenny S. Martínez, Horizontal Structuring, in THE OXFORD HANDBOOK OF COMPARATIVE CONSTITUTIONAL LAW 548, 572 (Michel Rosenfeld & András Sajó eds., 2012).
252 Sweet, supra note 86, at 824.
retains the option to overrule decisions, then this is a reasonable compromise. As Stone Sweet says of the constitutional courts of Europe that are more politicized but also more actively engaged in the protection of rights, “[t]he erosion of traditional separation of powers notions is the tax we pay for these benefits [of greater rights protection].”

Fourth, a concession to realism: should the FOB make a ruling that is fundamentally damaging to Facebook or out-of-step with user values, Facebook is likely to step in and overrule it. Indeed, it may even have a duty to its shareholders to do so. It is better to accept this reality from the outset, instead of creating a system that will only be undermined. This pragmatism will create a more robust and durable institution in the long run.

Finally, weak-form review is more suitable to the nature of the online speech ecosystem. Facebook is one part of a rapidly expanding and dynamic global information environment, where norms of communication are changing on screens before our eyes and global society is still reckoning with the consequences of the hyperconnected information glut of the modern era. Weak-form review better facilitates the capacity to evolve and revise. Indeed, it seems the only realistic course. Entrenching decisions or making it too costly to take into account ongoing changes would cause FOB decisions to endure past their relevance. As Kadri acknowledges, “it’s true that Facebook has good reason to want some flexibility to alter its speech policies, as it constantly learns and adapts to tackle the Sisyphean task of satisfying a “community” of more than 2 billion people. Binding itself to a 2018 version of its rules will surely be untenable. This is a fundamental point that needs to inform the entire thinking around institutions responsible with developing norms for online speech. There is still so much unknown about the new online speech environment, and research that needs to be done. Despite their dominance, these platforms are still a relatively young phenomenon. Platform rules should be able to be informed by new revelations. Facebook’s recent “pivot to privacy,” for example, shows how both platforms and users can rethink their fundamental priorities and values as the ramifications of the new online speech ecosystem become apparent. Governance structures need to be able to facilitate and respond to these developments.

Of course, the FOB can and should itself revise its decisions by distinguishing or updating precedent on the basis of “changed circumstances,” new evidence, or developing norms. But this is an inadequate solution because the FOB cannot update decisions proactively – a review body is necessarily responsive to claims and disputes brought to it. An example is the updates Facebook made to its approach to violent extremism in light of the Christchurch shooting. Facebook’s policies were being updated by the day in light of its experience trying to stop the spread of a live video of the attack. It is now experimenting with audio-based technology to identify designated videos, as well as sharing URLs with other members of industry. Facebook has also revised its response to “white nationalism” and “white separatism,” as part of its acknowledgment that it needs “to get better and faster at finding and removing hate from our platforms.” Facebook will begin removing such content, and also start directing people who post it to a nonprofit dedicated to helping people leave hate groups. Before these changes, Facebook did not prohibit

254 Sweet, supra note 86, at 828.
255 Tushnet, supra note 88, at 450 n309.
256 Kadri, supra note 228.
259 Id.
this kind of content on the grounds that concepts of nationalism and separatism are “an important part of people’s identity.”261 These are difficult issues, and there are no doubt people who would argue that Facebook has made the wrong call and such speech should not be censored.262 But the Christchurch shooting made clear the number of people interested in spreading this kind of content, and in a way that circumvents normal platform rules.263 As the President of Microsoft acknowledged, “it’s clear that we need to learn from and take new action based on what happened in Christchurch.”264 Events such as these can help inform long-standing debates about the merits of censorship versus counter-speech, which need to evolve to take account of the new empirical realities of how speech works online.265

FOB decisions should not stand in the way of such evolutions. It is easy to imagine a situation where the FOB had issued a ruling in different circumstances that would constrain Facebook’s response to the Christchurch shooting: for example, a holding that sharing a URL could never be “hate speech” under the Community Standards, or that the value of “voice” prevented Facebook from censoring URL sharing. Or there may be cases where the FOB issues a decision based on an understanding of how content such as hate speech works in the offline world, which is subsequently shown to not hold in the online ecosystem. As Scanlon urges, because rights contain “a significant empirical component, our understanding of a right can always be upset by evidence that forces a change in these empirical beliefs.”266 Again, such FOB decisions do not become worthless when Facebook seeks to reverse them – they will force a public dialogue about why Facebook is changing its rules or priorities.

There are many other examples of assumptions or norms that have been displaced or proven false as more is learnt about speech online. Researchers have been updating their findings on whether there is a so-called “backfire effect” to labelling something as untrue.267 This has important ramifications for the best way to treat misinformation, an issue which platforms have been grappling with since the “Fake News” scandals of 2016. Early findings about whether platforms increase “echo chambers” is also being disputed.268 Facebook can, and should, do much more to facilitate this kind of research by being more open with its data.269 But the company should not then be handicapped in its capacity to act on new findings.

It is the strong-form nature of review in the US that has lead scholars such as Tim Wu to argue that First Amendment jurisprudence has been made irrelevant by the changes technology has wrought to the speech environment.270 Changing First Amendment jurisprudence is a heavy lift, precisely because of the

261 Id.
262 See, eg, NADINE STROSSEN, HATE: WHY WE SHOULD RESIST IT WITH FREE SPEECH, NOT CENSORSHIP (Inalienable rights series, 2018).
265 Research is being done to better understand these dynamics: Susan Benesch, Launching today: new collaborative study to diminish abuse on Twitter, Medium (Apr. 712, 2018), https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc.
268 Andrew Guess, Avoiding the echo chamber about echo chambers, Trust, Media and Democracy (Feb. 903, 2018), https://medium.com/trust-media-and-democracy/avoiding-the-echo-chamber-about-echo-chambers-6e1f1a1a0f39.
269 Lazer et al., supra note 180.
270 Tim Wu, Is the First Amendment Obsolete?, Knight First Amendment Institute’s Emerging Threats (2017), https://knightcolumbia.org/sites/default/files/content/Emerging%20Threats%20Tim%20Wu%20Is%20The%20First
difficulty of amending the US Constitution or overturning Supreme Court decisions. Given how much is still unknown about online speech, and how this impacts offline lives, humility about the finality of the FOB’s decisions is warranted.

3. A Check on Users Too

There is one area, however, in which strong-form review’s ability to ossify the shape of rights could be a significant benefit: protection of minorities. Courts are generally an important mechanism for the protection of minority rights. This is one of the key reasons why courts are insulated from ordinary political processes, so that they are not constrained from playing this counter-majoritarian role in necessary cases. Indeed, Dworkin argued that the essence of constitutionalism is “the theory that the majority must be restrained to protect individual rights.” Such issues often arise in the context of content moderation, where many of the biggest content moderation controversies have been about concerns for the rights of minorities. As the UN Special Rapporteur on Freedom of Expression has noted, “[t]he vagueness of hate speech and harassment policies has triggered complaints of inconsistent policy enforcement that penalizes minorities while reinforcing the status of dominant or powerful groups. The FOB could have an especially important role in protecting the speech of those who are otherwise likely to be drowned out on the platform or who are underrepresented in the communities of engineers and content moderators who control speech on Facebook.

Zuckerberg writes in his Blueprint that “Just as our board of directors is accountable to our shareholders, [the FOB] would be focused only on our community.” He is right to distinguish between focus on the user community and accountability to it. While it is important that the FOB works to further the welfare of users (as distinguished from the welfare of Facebook as a company), this is not the same thing as making decisions that simply reflect user preferences. Indeed, “[a]n independent judiciary does not take a poll before rendering its decisions.” But there is the possibility that the FOB’s protective role is undermined if Facebook simply has the capacity to overrule FOB decisions that are unpopular with majorities. I do not believe, however, that this risk outweighs the potential benefit of weak-form review, and the need for content moderation rules to be able to be dynamically updated. It is true that there is little to guard against Facebook overruling the FOB in the name of majoritarianism beyond Facebook’s good faith implementation of the FOB’s independent oversight and a commitment to “rule of law” style principles such that it will not overturn decisions simply on the basis that they are unpopular. But my argument has been that the very fact that content moderation decisions are controversial and will tend to upset significant portions of users is part of the reason Facebook is setting the FOB up. This desire to

---

271 RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 175 (1977). See also, e.g., THE FEDERALIST NO. 78 (Hamilton) (“This independence of the judges is equally requisite to guard the Constitution and the rights of individuals from the effects of those ill humors, which the arts of designing men, or the influence of particular conjunctures, sometimes disseminate among the people themselves, and which, though they speedily give place to better information, and more deliberate reflection, have a tendency, in the meantime, to occasion dangerous innovations in the government, and serious oppressions of the minor party in the community.”); West Virginia State Board of Education v. Barnette, 638 (Justice Jackson writing that one’s “right to life, liberty and property, to free speech, free press, freedom of worship and assembly, and other fundamental rights” may not be submitted to vote.”)

272 See, e.g., Schechter, supra note 137 (reporting on a letter from civil rights organizations, including the ACLU, criticizing Facebook: “Activists in the Movement for Black Lives have routinely reported the takedown of images discussing racism and during protests, with the justification that it violates Facebook’s Community Standards. At the same time, harassment and threats directed at activists based on their race, religion, and sexual orientation is thriving on Facebook.”); Julia Angwin, supra note 47.

273 Kaye, supra note 13, at 10.

274 Zuckerberg, supra note 1.

275 Cross, supra note 130, at 559.
outsource controversy suggests that mere unpopularity of a decision would not be a reason for Facebook to intervene because then it would be undermining its own project to distance itself from the controversies. In this sense, the lack of democratic accountability or elections of Facebook’s management could provide further protection from simple majoritarianism.

B. Public Reasoning

Public reason-giving is the defining characteristic of the FOB. Facebook already has internal teams that continually review content moderation standards and consult outside experts.276 The utility of the FOB is not merely in having experts review Facebook’s decisions and coming to a “right” outcome. If it were, the internal processes would be sufficient. Instead, the FOB is intended to serve a function separate and additional to having experts weigh in on Facebook’s rules. The core value offering of the FOB is the issuing of public explanations for rulings.277 As Noah Feldman wrote in his original white paper proposing the FOB:

The advantage enjoyed by real-life constitutional courts is that they openly address difficult cases, and so derive credit and legitimacy from being principled. They make mistakes, and correct them. Their rules evolve with changing technology and ideas. And instead of blaming them for this, we mostly validate their efforts. … Right now, the platforms are already doing plenty of balancing work. But they aren’t doing it transparently or in a legal-logical way. Changes are greeted with outrage rather than respectful engagement. All that could change if the platforms provided a forum for argument, openly considered opposing views, and announced the reasoning behind their decisions on a case-by-case basis.278

1. The Purpose of Public Reasons

What exactly, then, is the purpose of public reasoning? An obvious answer might be that in an individual case adequate reasons are an incident of due process and fairness, by allowing a user to know the reasons for a decision (especially an adverse one) involving their freedom of expression on the platform.279 This is why many common law jurisdictions have a judicial duty to give reasons as an incident of due process280 and all constitutional courts of Europe are obliged to give reasons for their decisions.281 This is also why the Santa Clara Principles, a set of minimum requirements for transparency and accountability for content

---

276 Facts About Content Review on Facebook, Facebook Newsroom (Dec. 28, 2018), https://newsroom.fb.com/news/2018/12/content-review-facts/ (Facebook noting it holds “a global forum held every two weeks where we discuss potential changes to our policies. It includes experts from around the world with deep knowledge of relevant laws, online safety, counter-terrorism, operations, public policy, communications, product, and diversity.”).
278 Darmé et al., supra note 133, at 138.
279 See, eg, Court of Appeal in Flannery v Halifax Estate Agencies Ltd [2000] 1 WLR 377, 381 (“The duty is a function of due process, and therefore of justice. … [F]airness surely requires that the parties especially the losing party should be left in no doubt why they have won or lost. … [A] requirement to give reasons concentrates the mind; if it is fulfilled, the resulting decision is much more likely to be soundly based on the evidence than if it is not.”).
281 SWEET, supra note 8, at 56 n15.
moderation proposed by a group of leading civil society organizations, academics and experts, requires that users be given “a statement of reasoning sufficient to allow the user to understand the decision.” But, as has been noted repeatedly throughout this article, the due process offerings of the FOB will be limited due to the small fraction of cases it will adjudicate. Furthermore, it would be possible for affected users to be given decisions without them being made public, which would also diminish privacy concerns in releasing public opinions. Therefore, there must be other, more general rationales for public reason-giving.

As a practical matter, giving reasons assists in creating a consistent body of case law by allowing the FOB and users to understand the basis on which a particular decision is made and distinguish or apply it in future cases. Stone Sweet describes giving defensible reasons that are accepted as a precedential interpretation of constitutional meaning as “necessary conditions for the emergence of effective review systems.” But reasons do not need to be available to the public at large for moderators to apply them. So the goal of reason giving is still more general than that.

Fundamentally, the FOB is a response to calls for decisional transparency and a “global dialogue” with users and stakeholders about the impacts and justifications of Facebook’s rules. Platforms generally have been increasingly criticized as having opaque decision-making that interferes with their obligations of clarity, specificity and predictability. Public reasoning, then, is about facilitating this dialogue and legitimizing Facebook’s exercise of power over its users. As Rawls argued, in a pluralistic society where there will always be disagreement about what rule is best, the exercise of power over those who disagree with decisions is only legitimized through public reasoning that proceeds in a way people might be expected to respect. As Fallon notes, when decision-makers rely on reasons that reasonable people would acknowledge as fair (and not, for example, idiosyncratic or partisan), this creates legitimacy even if people might reach different ultimate judgments. Therefore, the goal is emphatically not to create collective agreement, but to allow for reasoned disagreement. Tom Tyler’s research has shown that people’s judgments of legitimacy do not depend primarily on their obtaining favorable outcomes, but are more strongly influenced by the processes and procedures authorities use, including whether they afford participation, demonstrate impartiality and show respect for people’s interests as worthy of consideration. Therefore, public reasoning is crucial to the FOB’s central goal of creating legitimacy for Facebook’s rules.

Until now, Facebook has exercised substantial power over not only individual speech but entire societies’ public discourse while giving little justification for the way in which it does so. The FOB is a step towards changing this. Just as constitutional courts can provide “a focal point for a new rhetoric of state legitimacy” during democratic transitions, so too might the FOB provide a forum for working through the principles in this new era of online governance.

Two opposing examples are illustrative of the benefits of public reasoning. In the first example, a lack of public reasoning delegitimized a decision even for those that agreed with it. In the second, reasoning

---

283 Sweet, supra note 86, at 825.
284 Ash et al., supra note 15, at 13; Kaye, supra note 13.
285 Kaye, supra note 13.
286 RAWLS, supra note 26, at 217; see also FALLO, supra note 72, at 12; Lawrence B. Solum, Procedural Justice, 78 S. CAL. L. REV. 181, 230 (2004).
287 FALLO, supra note 72, at 128.
288 Tom R. Tyler, Procedural Justice, Legitimacy, and the Effective Rule of Law, 3 CRIME AND JUSTICE 283 (2003); Facebook Data Transparency Advisory Group, supra note 84, at 34.
289 Sweet, supra note 86, at 827.
helped the affected user come to accept a decision they still thought was wrong.

The first example is the controversy over Alex Jones’ presence on Facebook. Leaked emails show the difficulty internal executives were having deciding whether a particular post violated Facebook’s Community Standards.290 In one email, an executive referred to the fact that the number of comments on the post that violated the Community Standards “did not meet the threshold for deletion.”291 UK executives then pointed out local context, noting that the image is famous in the UK and “widely acknowledged to be anti-Semitic” there, justifying deletion.292 These are the kinds of arguments the FOB could help ventilate. The “threshold” referred to is not defined in any public document. Whether or not this is a legitimate way of determining whether content should be removed, the fact that it is applied in a non-transparent way by executives that seem to disagree about whether the standard is met deprives it of legitimacy as a rule that users can reasonably agree to be bound by. The FOB might ultimately decide to apply the same standard – again, the benefits of the FOB are not necessarily to reach different or “better” outcomes. But if the FOB did decide to apply this test, it would need to explain this decision publicly which would allow for greater consistency in its application as well as contestation about the appropriateness of the threshold in the first place. Similarly, public reasoning can excavate local context to inform users unfamiliar with it why a post has special meaning in different circumstances. But the importance of public reasoning is most evident in the reaction to the ultimate decision to take down his pages altogether. Although many commentators had been calling for his removal for some time, the response was mixed. Few defended Jones, but there was frustration with the way Facebook executed the ban and the lack of transparency around the reason or timing.293 Without public reasoning, even those that agreed with the decision thought it was illegitimate.

The second example is from a different platform but illustrates the benefits of giving understandable rationales. David Neiwert’s Twitter account was suspended when he changed his profile picture to the cover of his book about the alt-right, which included KKK hoods.294 Neiwert thought the suspension was wrong and refused to change the picture. From his perspective, the image was about analyzing hate, not promoting it. Representatives from Twitter reached out and explained that he was making the distinction he was making, the company takes a no-tolerance stance on such images in profile pictures.

291 Id. Infringing comments amounted to about 4% of the total number of comments on the post.
292 Id.
293 April Glaser, Why Facebook’s Latest Ban of Alex Jones and Company Was So Underwhelming, Slate Magazine ("Facebook has the power to punish wrongdoers, as it did on Thursday. But we don’t know its full rationale for doing so, nor do we know who will be next."). (May 598, 2019), https://slate.com/technology/2019/05/facebook-alex-jones-ban-underwhelming.html; Bret Stephens, Opinion | Facebook’s Unintended Consequence, THE NEW YORK TIMES, May 896, 2019, at ("The deeper problem is the overwhelming concentration of technical, financial and moral power in the hands of people who lack the training, experience, wisdom, trustworthiness, humility and incentives to exercise that power responsibly. ... the decision to absolutely ban certain individuals will always be a human one. It will inevitably be subjective.")., https://www.nytimes.com/2019/05/03/opinion/facebook-free-speech.html; Why Facebook’s Bans Warrant Concern, National Review ("This means a person can potentially face social-media bans even if they comply with every syllable of the company’s speech rules on the company’s platform. The potential for abuse is obvious, as is the potential chilling effect") (May 0, 2019), https://www.nationalreview.com/corner/why-facebook-bans-warrant-concern/; Emily Stewart, Facebook bans Alex Jones, Infowars, Louis Farrakhan, and others it deems “dangerous,” Vox ("It’s not clear why Facebook is doing this now, but pressure for it to take action, and has been mounting for quite some time, and the decision is probably at least in part an effort to get some positive PR."). (May 2–2019), https://www.vox.com/recode/2019/5/2/18527357/facebook-bans-alex-jones-louis-farrakhan-infowars.
because they are more prominently displayed on the site. Neiwert wrote that the conversation “was cooperative and [they were] genuinely interested in my input. These Twitter officials were able to persuade me, at least, that they very much share my concerns.” Neiwert still disagreed with the decision, but once he understood the reasoning he agreed to change his profile picture and his account was reinstated.

These examples illustrate the ultimate hope for the FOB. As Feldman summarizes:

Some controversy and disagreement over the [FOB’s] decisions is inevitable. But even when it occurs, it will come in the context of the public understanding that Facebook is publicly and responsibly grappling with balancing values in cases that have no simple right answers. This repeated engagement should produce legitimacy for the decision process, and a new narrative for Facebook’s engagement with these problems.

Another benefit of public reasoning is its ability to help in norm-setting. The FOB can provide contestation and explanation of norms in a more public forum. This in turn might allow greater visibility and participation in content moderation decisions through public reasoning, the FOB can embed the process of rule formation in a broader community and help norms be formed and tested.

Public reasoning is also an important constraint on the FOB itself. Although it should be independent, this does not mean the FOB should be unconstrained. Giving principled reasons is the primary way that judges can be held accountable for their exercise of power. As Mark Tushnet writes, “the desideratum is not judicial independence alone but rather judicial independence coupled with accountability to law.” He explains that “without the latter, independent judges can act arbitrarily and so anticonstitutionally.

The FOB will need to develop a discourse that centered around norms that distinguishes it from a mere political institution. It is the norms and constraints of principled reason-giving that will reassure users that they are not merely “exchanging one set of tyrants for another.”

This legitimacy-creating constraint is especially when ambiguity is common. Accountability to law does not mean demonstration that the law requires a particular result. Indeed, as Rawls says, “public reason often allows more than one reasonable answer to any particular question.” As Roux has observed, “For a constitutional court successfully to establish and maintain its independence, after all, judges do not need to convince everybody all of the

295 Id.
296 Darmé et al., supra note 133, at 145.
299 Roderick A. Macdonald & Hoi Kong, Judicial Independence as a Constitutional Virtue, in The Oxford Handbook of Comparative Constitutional Law 831, 839 (Michel Rosenfeld & András Sajó eds., 2012).
300 Tushnet, supra note 88, at 419. See also Guarnieri & Pedertzoli, supra note 184, at 45; Cross, supra note 130, at 558.
301 Tushnet, supra note 130. n145. See also Guarnieri & Pedertzoli, supra note 184, at 196.
302 Guarnieri & Pedertzoli, supra note 184, at 196.
304 RAWLS, supra note 26, at 240.
time that their reasoning methods are determinate. All that they need to do is to respect the methods that have come to be seen by the legal community and the broader legal culture as legitimate.”305

2. The Challenge of Giving Acceptable Reasons

There are particular aspects of the FOB’s context that will make its use of public reasoning to create legitimacy uniquely challenging.

First, neither the FOB itself nor the community it speaks to are socialized in a particular style of discourse, such as legal reasoning. The FOB’s reasons and those that read them will not draw on what Llewellyn called law’s “steadying factors” which provides accepted doctrinal techniques for legal argument.306 Judicial legitimacy is itself at least partially rooted in adherence to this craft.307 Of course, some contest the constraining power of these norms and even their existence at all.308 But most accept that legal decisions are something different to mere politics.309 So while there is some degree of choice inherent in the process of construction, most constitutional scholars for example agree there is still something distinctively legal inherent in the process of constitutional construction.310 Setting aside the fundamental debate over legal realism, it is sufficient for present purposes to note that finding a mode of discourse that can be accepted as legitimizing in Facebook’s vast and diverse community will only be harder when not tied to a previously existing set of professional norms.

This is exacerbated because, second, there are no accepted global norms to guide the substantive issues. Discussing the diversity of views on freedom of expression, Facebook’s Head of Global Policy Management has written:311

Laws differ from country to country, not only in their text, but also in how their enforcement is pursued by authorities and how they are applied by courts. Even within the confines of one country, and even within common law jurisdictions, courts differ in their interpretation of laws. To complicate matters, laws evolve. New laws are passed, old laws are amended, and courts invalidate some laws altogether. … [E]ven assuming one government’s laws represent the social norms of its people, those laws will not represent the norms of every person on a global social media service.

Adopting international law standards does not simplify the matter. These are not universally endorsed.312 Nor are they themselves stable – for example, “international doctrine and practice relating to prohibition of hate speech remain uneven.”313 Procedural elements of resolving disagreement are not so easily divorced from the underlying substance.314 Finding arguments that can reasonably be regarded as

306 KARL N LLEWELLYN, THE COMMON LAW TRADITION: DECIDING APPEALS 5, 19-23 (1960). Ginsburg notes that professional norms of fidelity to law are a kind of internalized ideology that help reduce the agency costs of using judges as agents to monitor administrative structures: Ginsburg, supra note 114, at 65.
307 Fallon, supra note 200, at 1826.
308 Id. at 1826–27.
309 FALLON, supra note 72, at 2.
310 Landau & Dixon, supra note 77, at 23.
313 Cleveland, supra note 55, at 225.
acceptable across such diverse substantive traditions will be especially challenging.

Third, compounding the difficulties, Facebook does not have an underlying rationale for free speech on its platform to justify its choices. As Adrienne Stone notes, even in nation states, “identifying the value or set of values underlying any single constitutional system of freedom of expression is likely to be difficult.”315 For Facebook, it is even more difficult. What is the purpose of speech on Facebook? The simple answer that it generates revenue by capturing attention is not one that will help satisfactorily justify outcomes in difficult cases. But at least the three most common rationales from other traditions do not fit easily in the context of Facebook: facilitating self-government; the search for truth; or respect for individual autonomy. This idea that freedom of expression is valuable because of its capacity to promote democratic self-government is the most widely adopted in modern legal systems,316 but seems inappropriate in the context of Facebook. Facebook is not a democracy—it is a business.317 It does not rely on popular will in setting its rules (a brief experiment with a limited such system in 2009 failed due to poor voting levels318). It has expressly disavowed any conception of its platform as facilitating the search for truth.319 In any event, its algorithm manipulates the “marketplace of ideas” which is metaphorically said to lead to truth’s revelation. This leaves the third dominant justification for freedom of speech: that based on individual autonomy.320 Perhaps this is the most fitting understanding of Facebook’s justification for free speech, but still sits uneasily with the fact that Facebook regularly censors speech. It is therefore difficult to pin down any underlying theory behind Facebook’s content moderation system which could inform the FOB’s work and which could satisfy those that disagree with the FOB’s decisions that they are based on some more fundamental principle. As discussed above, Facebook’s statement of values might mitigate this, but only to a certain extent. A decision to prioritize “safety” would give a clear guiding principle in many individual decisions. But if Facebook decides its central value is “voice”, this will not necessarily make trade-offs easier. Whose voice should be given priority, for example? Justifying decisions to prioritize voice without a thesis of the underlying purpose of voice will limit the extent to which justifications can legitimize choices.

These are the challenges for the FOB fulfilling its role as an “exemplar of public reason.”321 Nevertheless, it is a central and important goal for the institution. As more expression occurs in these online spaces, it is unavoidable that there will need to be some body that performs this role of rationalizing contested and contestable decisions about how to regulate important public discourse. For now, the FOB is the first attempt at such an institution. It might be a canary in the coal mine of the future of online governance. No doubt, much will be learnt from its successes and, perhaps inevitably, blunders.

316 Id. at 414.
317 Pozen, supra note 30 (explaining that there is a “tension between Facebook’s seemingly sincere concern for free speech values and its explicit aspiration to make users feel socially safe and ‘connected’ [and hence to maximize the time they spend on the site], a tension that is shaped by market forces but ultimately resolved by benevolent leader and controlling shareholder Zuckerberg.”).
320 Stone, supra note 315, at 413–14; FREDERICK F. SCHAUER, FREE SPEECH: A PHILOSOPHICAL ENQUIRY (1982).
321 RAWLS, supra note 26, at 231.
VI. CONCLUSION

The FOB represents an important innovation, and a welcome attempt to disperse the enormous power over online discourse held by Facebook. More importantly, however, it will serve to make that power more transparent and legitimate by facilitating dialogues around how and why Facebook’s power is exercised in the first place. This is a more modest goal than becoming an independent source of universally accepted free speech norms, but it is still incredibly ambitious for an institution that is breaking new ground. As Stone Sweet writes about constitutional courts in regime transitions:

The ultimate measure of legitimacy for any [constitutional court] may well be its success at helping the polity construct a new ‘constitutional identity’—a massive undertaking. … [A]s Schepple writes, a [constitutional court] is often ‘the primary mechanism’ for organizing the transition away from the former ‘regime of horror’ to constitutional democracy. Insofar as [constitutional courts] are successful, the legitimacy of the constitution, as a basic framework for the exercise of public authority, will become indistinguishable from the regime's political legitimacy.

For the best chances of success, Facebook and the FOB should be humble and acknowledge the very real limitations that such a body will face. As a Facebook representative notes, “We are very much at the beginning of this process—it has not been done before.”322 There is no true model to base the new body on. Deciding what the FOB will be includes accepting what it cannot be—this includes acknowledging that it cannot be a way to bring due process to any but the smallest fraction of content moderation decisions. Nor can it authoritatively resolve clashing ideas of freedom of expression. But, if done right, the FOB may be able to bring a greater sense of legitimacy and acceptance to Facebook’s content moderation ecosystem. This is a massive undertaking, and will require a healthy dose of “constitutional luck.”323 History teaches us that ultimately many variables for constitutional success are beyond the ability of designers to control and success or failure is necessarily contingent. This may be unsatisfying. But the question should not be whether the FOB is inevitably a perfect institution or an ideal-type of due process and transparency that will bind Facebook to stringent human rights standards. The question is whether it is better than the alternative: the current haphazard, opaque process that draws inspiration from Kafka more than Kelsen. By that standard, the FOB shows real promise.

322 Archibong, supra note 112.
New York state law requires cable operators to set aside channels on their cable systems for public access. Those channels are operated by the cable operator unless the local government chooses to itself operate the channels or designates a private entity to operate the channels. New York City (the City) has designated a private nonprofit corporation, petitioner Manhattan Neighborhood Network (MNN), to operate the public access channels on Time Warner’s cable system in Manhattan. Respondents DeeDee Halleck and Jesus Papoleto Melendez produced a film critical of MNN to be aired on MNN’s public access channels. MNN televised the film. MNN later suspended Halleck and Melendez from all MNN services and facilities. The producers sued, claiming that MNN violated their First Amendment free-speech rights when it restricted their access to the public access channels because of the content of their film. The District Court dismissed the claim on the ground that MNN is not a state actor and therefore is not subject to First Amendment constraints on its editorial discretion. Reversing in relevant part, the Second Circuit concluded that MNN is a state actor subject to First Amendment constraints.

Held: MNN is not a state actor subject to the First Amendment. Pp. 5–16.

(a) The Free Speech Clause of the First Amendment prohibits only governmental, not private, abridgment of speech. See, e.g., Denver Area Ed. Telecommunications Consortium, Inc. v. FCC, 518 U. S. 727, 737. This Court’s state-action doctrine distinguishes the government from individuals and private entities. Pp. 5–14.

(1) A private entity may qualify as a state actor when, as rele-
Syllabus

Venant here, the entity exercises “powers traditionally exclusively reserved to the State.” *Jackson v. Metropolitan Edison Co.*, 419 U. S. 345, 352. The Court has stressed that “very few” functions fall into that category. *Flagg Bros., Inc. v. Brooks*, 436 U. S. 149, 158. The relevant function in this case—operation of public access channels on a cable system—has not traditionally and exclusively been performed by government. Since the 1970s, a variety of private and public actors have operated public access channels. Early Manhattan public access channels were operated by private cable operators with some help from private nonprofit organizations. That practice continued until the early 1990s, when MNN began to operate the channels. Operating public access channels on a cable system is not a traditional, exclusive public function. Pp. 6–8.

(2) The producers contend that the relevant function here is more generally the operation of a public forum for speech, which, they claim, is a traditional, exclusive public function. But that analysis mistakenly ignores the threshold state-action question. Providing some kind of forum for speech is not an activity that only governmental entities have traditionally performed. Therefore, a private entity who provides a forum for speech is not transformed by that fact alone into a state actor. See *Hudgens v. NLRB*, 424 U. S. 507, 520–521. Pp. 8–10.

(3) The producers note that the City has designated MNN to operate the public access channels on Time Warner’s cable system, and that the State heavily regulates MNN with respect to those channels. But the City’s designation is analogous to a government license, a government contract, or a government-granted monopoly, none of which converts a private entity into a state actor—unless the private entity is performing a traditional, exclusive public function. See, e.g., *San Francisco Arts & Athletics, Inc. v. United States Olympic Comm.*, 483 U. S. 522, 543–544. And the fact that MNN is subject to the State’s extensive regulation “does not by itself convert its action into that of the State.” *Jackson*, 419 U. S., at 350. Pp. 11–14.

(b) The producers alternatively contend that the public access channels are actually the City’s property and that MNN is essentially managing government property on the City’s behalf. But the City does not own or lease the public access channels and does not possess any formal easement or other property interest in the channels. It does not matter that a provision in the franchise agreements between the City and Time Warner allowed the City to designate a private entity to operate the public access channels on Time Warner’s cable system. Nothing in the agreements suggests that the City possesses any property interest in the cable system or in the public access channels on that system. Pp. 14–15.
Syllabus

882 F. 3d 300, reversed in part and remanded.

KAVANAUGH, J., delivered the opinion of the Court, in which ROBERTS, C. J., and THOMAS, ALITO, and GORSUCH, JJ., joined. SOTOMAYOR, J., filed a dissenting opinion, in which GINSBURG, BREYER, and KAGAN, JJ., joined.
SUPREME COURT OF THE UNITED STATES  

No. 17–1702  

MANHATTAN COMMUNITY ACCESS CORPORATION,  
ET AL., PETITIONERS v. DEEDEE HALLECK, ET AL.  

ON WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE SECOND CIRCUIT  

[June 17, 2019]  

JUSTICE KAVANAUGH delivered the opinion of the Court.  

The Free Speech Clause of the First Amendment constrains governmental actors and protects private actors. To draw the line between governmental and private, this Court applies what is known as the state-action doctrine. Under that doctrine, as relevant here, a private entity may be considered a state actor when it exercises a function “traditionally exclusively reserved to the State.” Jackson v. Metropolitan Edison Co., 419 U. S. 345, 352 (1974).  

This state-action case concerns the public access channels on Time Warner’s cable system in Manhattan. Public access channels are available for private citizens to use. The public access channels on Time Warner’s cable system in Manhattan are operated by a private nonprofit corporation known as MNN. The question here is whether MNN—even though it is a private entity—nonetheless is a state actor when it operates the public access channels. In other words, is operation of public access channels on a cable system a traditional, exclusive public function? If so, then the First Amendment would restrict MNN’s exercise
of editorial discretion over the speech and speakers on the public access channels.

Under the state-action doctrine as it has been articulated and applied by our precedents, we conclude that operation of public access channels on a cable system is not a traditional, exclusive public function. Moreover, a private entity such as MNN who opens its property for speech by others is not transformed by that fact alone into a state actor. In operating the public access channels, MNN is a private actor, not a state actor, and MNN therefore is not subject to First Amendment constraints on its editorial discretion. We reverse in relevant part the judgment of the Second Circuit, and we remand the case for further proceedings consistent with this opinion.

I

A

Since the 1970s, public access channels have been a regular feature on cable television systems throughout the United States. In the 1970s, Federal Communications Commission regulations required certain cable operators to set aside channels on their cable systems for public access. In 1979, however, this Court ruled that the FCC lacked statutory authority to impose that mandate. See *FCC v. Midwest Video Corp.*, 440 U. S. 689 (1979). A few years later, Congress passed and President Reagan signed the Cable Communications Policy Act of 1984. 98 Stat. 2779. The Act authorized state and local governments to require cable operators to set aside channels on their cable systems for public access. 47 U. S. C. §531(b).

The New York State Public Service Commission regulates cable franchising in New York State and requires cable operators in the State to set aside channels on their cable systems for public access. 16 N. Y. Codes, Rules & Regs. §§895.1(f), 895.4(b) (2018). State law requires that use of the public access channels be free of charge and
Opinion of the Court

first-come, first-served. §§895.4(c)(4) and (6). Under state law, the cable operator operates the public access channels unless the local government in the area chooses to itself operate the channels or designates a private entity to operate the channels. §895.4(c)(1).

Time Warner (now known as Charter) operates a cable system in Manhattan. Under state law, Time Warner must set aside some channels on its cable system for public access. New York City (the City) has designated a private nonprofit corporation named Manhattan Neighborhood Network, commonly referred to as MNN, to operate Time Warner’s public access channels in Manhattan. This case involves a complaint against MNN regarding its management of the public access channels.

Because this case comes to us on a motion to dismiss, we accept the allegations in the complaint as true. See Ashcroft v. Iqbal, 556 U. S. 662, 678 (2009).

DeeDee Halleck and Jesus Papoleto Melendez produced public access programming in Manhattan. They made a film about MNN’s alleged neglect of the East Harlem community. Halleck submitted the film to MNN for airing on MNN’s public access channels, and MNN later televised the film. Afterwards, MNN fielded multiple complaints about the film’s content. In response, MNN temporarily suspended Halleck from using the public access channels.

Halleck and Melendez soon became embroiled in another dispute with MNN staff. In the wake of that dispute, MNN ultimately suspended Halleck and Melendez from all MNN services and facilities.

Halleck and Melendez then sued MNN, among other parties, in Federal District Court. The two producers claimed that MNN violated their First Amendment free-speech rights when MNN restricted their access to the
public access channels because of the content of their film.

MNN moved to dismiss the producers’ First Amendment claim on the ground that MNN is not a state actor and therefore is not subject to First Amendment restrictions on its editorial discretion. The District Court agreed with MNN and dismissed the producers’ First Amendment claim.

The Second Circuit reversed in relevant part. 882 F. 3d 300, 308 (2018). In the majority opinion authored by Judge Newman and joined by Judge Lohier, the court stated that the public access channels in Manhattan are a public forum for purposes of the First Amendment. Reasoning that “public forums are usually operated by governments,” the court concluded that MNN is a state actor subject to First Amendment constraints. Id., at 306–307. Judge Lohier added a concurring opinion, explaining that MNN also qualifies as a state actor for the independent reason that “New York City delegated to MNN the traditionally public function of administering and regulating speech in the public forum of Manhattan’s public access channels.” Id., at 309.

Judge Jacobs dissented in relevant part, opining that MNN is not a state actor. He reasoned that a private entity’s operation of an open forum for speakers does not render the host entity a state actor. Judge Jacobs further stated that the operation of public access channels is not a traditional, exclusive public function.

We granted certiorari to resolve disagreement among the Courts of Appeals on the question whether private operators of public access cable channels are state actors subject to the First Amendment. 586 U. S. __ (2018). Compare 882 F. 3d 300 (case below), with Wilcher v. Akron, 498 F. 3d 516 (CA6 2007); and Alliance for Community Media v. FCC, 56 F. 3d 105 (CADC 1995).
Ratified in 1791, the First Amendment provides in relevant part that “Congress shall make no law . . . abridging the freedom of speech.” Ratified in 1868, the Fourteenth Amendment makes the First Amendment's Free Speech Clause applicable against the States: “No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law . . . .” §1. The text and original meaning of those Amendments, as well as this Court's longstanding precedents, establish that the Free Speech Clause prohibits only governmental abridgment of speech. The Free Speech Clause does not prohibit private abridgment of speech. See, e.g., Denver Area Ed. Telecommunications Consortium, Inc. v. FCC, 518 U.S. 727, 737 (1996) (plurality opinion); Hurley v. Irish-American Gay, Lesbian and Bisexual Group of Boston, Inc., 515 U.S. 557, 566 (1995); Hudgens v. NLRB, 424 U.S. 507, 513 (1976); cf. Miami Herald Publishing Co. v. Tornillo, 418 U.S. 241, 256 (1974).

In accord with the text and structure of the Constitution, this Court's state-action doctrine distinguishes the government from individuals and private entities. See Brentwood Academy v. Tennessee Secondary School Athletic Assn., 531 U.S. 288, 295–296 (2001). By enforcing that constitutional boundary between the governmental and the private, the state-action doctrine protects a robust sphere of individual liberty.

Here, the producers claim that MNN, a private entity, restricted their access to MNN's public access channels because of the content of the producers' film. The producers have advanced a First Amendment claim against MNN. The threshold problem with that First Amendment claim is a fundamental one: MNN is a private entity.

Relying on this Court's state-action precedents, the
producers assert that MNN is nonetheless a state actor subject to First Amendment constraints on its editorial discretion. Under this Court’s cases, a private entity can qualify as a state actor in a few limited circumstances—including, for example, (i) when the private entity performs a traditional, exclusive public function, see, e.g., *Jackson*, 419 U. S., at 352–354; (ii) when the government compels the private entity to take a particular action, see, e.g., *Blum v. Yaretsky*, 457 U. S. 991, 1004–1005 (1982); or (iii) when the government acts jointly with the private entity, see, e.g., *Lugar v. Edmondson Oil Co.*, 457 U. S. 922, 941–942 (1982).

The producers’ primary argument here falls into the first category: The producers contend that MNN exercises a traditional, exclusive public function when it operates the public access channels on Time Warner’s cable system in Manhattan. We disagree.

A

Under the Court’s cases, a private entity may qualify as a state actor when it exercises “powers traditionally exclusively reserved to the State.” *Jackson*, 419 U. S., at 352. It is not enough that the federal, state, or local government exercised the function in the past, or still does. And it is not enough that the function serves the public good or the public interest in some way. Rather, to qualify as a traditional, exclusive public function within the meaning of our state-action precedents, the government must have traditionally *and* exclusively performed the function. See *Rendell-Baker v. Kohn*, 457 U. S. 830, 842 (1982); *Jackson*, 419 U. S., at 352–353; *Evans v. Newton*, 382 U. S. 296, 300 (1966).

The Court has stressed that “very few” functions fall into that category. *Flagg Bros., Inc. v. Brooks*, 436 U. S. 149, 158 (1978). Under the Court’s cases, those functions include, for example, running elections and operating a

The relevant function in this case is operation of public access channels on a cable system. That function has not traditionally and exclusively been performed by government.

Since the 1970s, when public access channels became a regular feature on cable systems, a variety of private and public actors have operated public access channels, includ-

\(^1\) Relatedly, this Court has recognized that a private entity may, under certain circumstances, be deemed a state actor when the government has outsourced one of its constitutional obligations to a private entity. In West v. Atkins, for example, the State was constitutionally obligated to provide medical care to prison inmates. 487 U. S. 42, 56 (1988). That scenario is not present here because the government has no such obligation to operate public access channels.
Opinion of the Court

private cable operators; private nonprofit organizations; municipalities; and other public and private community organizations such as churches, schools, and libraries. See Denver Area, 518 U. S., at 761–762 (plurality opinion); R. Oringel & S. Buske, The Access Manager’s Handbook: A Guide for Managing Community Television 14–17 (1987).

The history of public access channels in Manhattan further illustrates the point. In 1971, public access channels first started operating in Manhattan. See D. Brenner, M. Price, & M. Meyerson, Cable Television and Other Nonbroadcast Video §6:29, p. 6–47 (2018). Those early Manhattan public access channels were operated in large part by private cable operators, with some help from private nonprofit organizations. See G. Gillespie, Public Access Cable Television in the United States and Canada 37–38 (1975); Janes, History and Structure of Public Access Television, 39 J. Film & Video, No. 3, pp. 15–17 (1987). Those private cable operators continued to operate the public access channels until the early 1990s, when MNN (also a private entity) began to operate the public access channels.

In short, operating public access channels on a cable system is not a traditional, exclusive public function within the meaning of this Court’s cases.

To avoid that conclusion, the producers widen the lens and contend that the relevant function here is not simply the operation of public access channels on a cable system, but rather is more generally the operation of a public forum for speech. And according to the producers, operation of a public forum for speech is a traditional, exclusive public function.

That analysis mistakenly ignores the threshold state-action question. When the government provides a forum
for speech (known as a public forum), the government may be constrained by the First Amendment, meaning that the government ordinarily may not exclude speech or speakers from the forum on the basis of viewpoint, or sometimes even on the basis of content. See, e.g., Southeastern Promotions, Ltd. v. Conrad, 420 U. S. 546, 547, 555 (1975) (private theater leased to the city); Police Dept. of Chicago v. Mosley, 408 U. S. 92, 93, 96 (1972) (sidewalks); Hague v. Committee for Industrial Organization, 307 U. S. 496, 515–516 (1939) (streets and parks).

By contrast, when a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment because the private entity is not a state actor. The private entity may thus exercise editorial discretion over the speech and speakers in the forum. This Court so ruled in its 1976 decision in Hudgens v. NLRB. There, the Court held that a shopping center owner is not a state actor subject to First Amendment requirements such as the public forum doctrine. 424 U. S., at 520–521; see also Lloyd Corp. v. Tanner, 407 U. S. 551, 569–570 (1972); Central Hardware Co. v. NLRB, 407 U. S. 539, 547 (1972); Alliance for Community Media, 56 F. 3d, at 121–123.

The Hudgens decision reflects a commonsense principle: Providing some kind of forum for speech is not an activity that only governmental entities have traditionally performed. Therefore, a private entity who provides a forum for speech is not transformed by that fact alone into a state actor. After all, private property owners and private lessees often open their property for speech. Grocery stores put up community bulletin boards. Comedy clubs host open mic nights. As Judge Jacobs persuasively explained, it “is not at all a near-exclusive function of the state to provide the forums for public expression, politics, information, or entertainment.” 882 F. 3d, at 311 (opinion concurring in part and dissenting in part).
In short, merely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints.

If the rule were otherwise, all private property owners and private lessees who open their property for speech would be subject to First Amendment constraints and would lose the ability to exercise what they deem to be appropriate editorial discretion within that open forum. Private property owners and private lessees would face the unappetizing choice of allowing all comers or closing the platform altogether. “The Constitution by no means requires such an attenuated doctrine of dedication of private property to public use.” Hudgens, 424 U. S., at 519 (internal quotation marks omitted). Benjamin Franklin did not have to operate his newspaper as “a stagecoach, with seats for everyone.” F. Mott, American Journalism 55 (3d ed. 1962). That principle still holds true. As the Court said in Hudgens, to hold that private property owners providing a forum for speech are constrained by the First Amendment would be “to create a court-made law wholly disregarding the constitutional basis on which private ownership of property rests in this country.” 424 U. S., at 517 (internal quotation marks omitted). The Constitution does not disable private property owners and private lessees from exercising editorial discretion over speech and speakers on their property.2

The producers here are seeking in effect to circumvent this Court’s case law, including Hudgens. But Hudgens is sound, and we therefore reaffirm our holding in that case.3

---

2A distinct question not raised here is the degree to which the First Amendment protects private entities such as Time Warner or MNN from government legislation or regulation requiring those private entities to open their property for speech by others. Cf. Turner Broadcasting System, Inc. v. FCC, 512 U. S. 622, 636–637 (1994).

3In Cornelius v. NAACP Legal Defense & Educational Fund, Inc.,
Next, the producers retort that this case differs from \textit{Hudgens} because New York City has designated MNN to operate the public access channels on Time Warner’s cable system, and because New York State heavily regulates MNN with respect to the public access channels. Under this Court’s cases, however, those facts do not establish that MNN is a state actor.

New York City’s designation of MNN to operate the public access channels is analogous to a government license, a government contract, or a government-granted monopoly. But as the Court has long held, the fact that the government licenses, contracts with, or grants a monopoly to a private entity does not convert the private entity into a state actor—unless the private entity is performing a traditional, exclusive public function. See, \textit{e.g.}, \textit{San Francisco Arts & Athletics}, 483 U. S., at 543–544 (exclusive-use rights and corporate charters); \textit{Blum}, 457 U. S., at 1011 (licenses); \textit{Rendell-Baker}, 457 U. S., at 840–841 (contracts); \textit{Polk County}, 454 U. S., at 319, n. 9, and 320–322 (law licenses); \textit{Jackson}, 419 U. S., at 351–352 (electric monopolies); \textit{Columbia Broadcasting System, Inc. v. Democratic National Committee}, 412 U. S. at 94, 120–121 (broadcast licenses); \textit{Moose Lodge No. 107 v. Irvis}, 407 U. S. 163, 176–177 (1972) (liquor licenses); cf. \textit{Trustees}.

this Court said in passing dicta that “a speaker must seek access to public property or to private property dedicated to public use to evoke First Amendment concerns.” 473 U. S. 788, 801 (1985). But \textit{Cornelius} dealt with government-owned property. As JUSTICE THOMAS explained in \textit{Denver Area Educational Telecommunications Consortium, Inc. v. FCC}, the Court’s admittedly imprecise and overbroad phrase in \textit{Cornelius} is not consistent with this Court’s case law and should not be read to suggest that private property owners or private lessees are subject to First Amendment constraints whenever they dedicate their private property to public use or otherwise open their property for speech. 518 U. S. 727, 827–828 (1996) (opinion concurring in judgment in part and dissenting in part).
MANHATTAN COMMUNITY ACCESS CORP. v. HALLECK

Opinion of the Court


Numerous private entities in America obtain government licenses, government contracts, or government-granted monopolies. If those facts sufficed to transform a private entity into a state actor, a large swath of private entities in America would suddenly be turned into state actors and be subject to a variety of constitutional constraints on their activities. As this Court’s many state-action cases amply demonstrate, that is not the law. Here, therefore, the City’s designation of MNN to operate the public access channels on Time Warner’s cable system does not make MNN a state actor.

So, too, New York State’s extensive regulation of MNN’s operation of the public access channels does not make MNN a state actor. Under the State’s regulations, air time on the public access channels must be free, and programming must be aired on a first-come, first-served basis. Those regulations restrict MNN’s editorial discretion and in effect require MNN to operate almost like a common carrier. But under this Court’s cases, those restrictions do not render MNN a state actor.

In Jackson v. Metropolitan Edison Co., the leading case on point, the Court stated that the “fact that a business is subject to state regulation does not by itself convert its action into that of the State.” 419 U. S., at 350. In that case, the Court held that “a heavily regulated, privately owned utility, enjoying at least a partial monopoly in the providing of electrical service within its territory,” was not a state actor. Id., at 358. The Court explained that the “mere existence” of a “regulatory scheme”—even if “extensive and detailed”—did not render the utility a state actor. Id., at 350, and n. 7. Nor did it matter whether the State had authorized the utility to provide electric service to the
community, or whether the utility was the only entity providing electric service to much of that community.

This case closely parallels Jackson. Like the electric utility in Jackson, MNN is “a heavily regulated, privately owned” entity. Id., at 358. As in Jackson, the regulations do not transform the regulated private entity into a state actor.

Put simply, being regulated by the State does not make one a state actor. See Sullivan, 526 U. S., at 52; Blum, 457 U. S., at 1004; Rendell-Baker, 457 U. S., at 841–842; Jackson, 419 U. S., at 350; Moose Lodge, 407 U. S., at 176–177. As the Court’s cases have explained, the “being heavily regulated makes you a state actor” theory of state action is entirely circular and would significantly endanger individual liberty and private enterprise. The theory would be especially problematic in the speech context, because it could eviscerate certain private entities’ rights to exercise editorial control over speech and speakers on their properties or platforms. Not surprisingly, as JUSTICE THOMAS has pointed out, this Court has “never even hinted that regulatory control, and particularly direct regulatory control over a private entity’s First Amendment speech rights,” could justify subjecting the regulated private entity to the constraints of the First Amendment. Denver Area, 518 U. S., at 829 (opinion concurring in judgment in part and dissenting in part).

In sum, we conclude that MNN is not subject to First Amendment constraints on how it exercises its editorial discretion with respect to the public access channels. To be sure, MNN is subject to state-law constraints on its editorial discretion (assuming those state laws do not violate a federal statute or the Constitution). If MNN violates those state laws, or violates any applicable contracts, MNN could perhaps face state-law sanctions or liability of some kind. We of course take no position on any potential state-law questions. We simply conclude
that MNN, as a private actor, is not subject to First Amendment constraints on how it exercises editorial discretion over the speech and speakers on its public access channels.

III

Perhaps recognizing the problem with their argument that MNN is a state actor under ordinary state-action principles applicable to private entities and private property, the producers alternatively contend that the public access channels are actually the property of New York City, not the property of Time Warner or MNN. On this theory, the producers say (and the dissent agrees) that MNN is in essence simply managing government property on behalf of New York City.

The short answer to that argument is that the public access channels are not the property of New York City. Nothing in the record here suggests that a government (federal, state, or city) owns or leases either the cable system or the public access channels at issue here. Both Time Warner and MNN are private entities. Time Warner is the cable operator, and it owns its cable network, which contains the public access channels. MNN operates those public access channels with its own facilities and equipment. The City does not own or lease the public access channels, and the City does not possess a formal easement or other property interest in those channels. The franchise agreements between the City and Time Warner do not say that the City has any property interest in the public access channels. On the contrary, the franchise agreements expressly place the public access channels “under the jurisdiction” of MNN. App. 22. Moreover, the producers did not allege in their complaint that the City has a property interest in the channels. And the producers have not cited any basis in state law for such a conclusion. Put simply, the City does not have “any formal
Opinion of the Court

easement or other property interest in those channels.” *Denver Area*, 518 U. S., at 828 (opinion of THOMAS, J.).

It does not matter that a provision in the franchise agreements between the City and Time Warner allowed the City to designate a private entity to operate the public access channels on Time Warner's cable system. Time Warner still owns the cable system. And MNN still operates the public access channels. To reiterate, nothing in the franchise agreements suggests that the City possesses any property interest in Time Warner’s cable system, or in the public access channels on that system.

It is true that the City has allowed the cable operator, Time Warner, to lay cable along public rights-of-way in the City. But Time Warner’s access to public rights-of-way does not alter the state-action analysis. For Time Warner, as for other cable operators, access to public rights-of-way is essential to lay cable and construct a physical cable infrastructure. See *Turner Broadcasting System, Inc. v. FCC*, 512 U. S. 622, 628 (1994). But the same is true for utility providers, such as the electric utility in *Jackson*. Put simply, a private entity’s permission from government to use public rights-of-way does not render that private entity a state actor.

Having said all that, our point here should not be read too broadly. Under the laws in certain States, including New York, a local government may decide to itself operate the public access channels on a local cable system (as many local governments in New York State and around the country already do), or could take appropriate steps to obtain a property interest in the public access channels. Depending on the circumstances, the First Amendment might then constrain the local government’s operation of the public access channels. We decide only the case before us in light of the record before us.
Opinion of the Court

*   *   *

It is sometimes said that the bigger the government, the smaller the individual. Consistent with the text of the Constitution, the state-action doctrine enforces a critical boundary between the government and the individual, and thereby protects a robust sphere of individual liberty. Expanding the state-action doctrine beyond its traditional boundaries would expand governmental control while restricting individual liberty and private enterprise. We decline to do so in this case.

MNN is a private entity that operates public access channels on a cable system. Operating public access channels on a cable system is not a traditional, exclusive public function. A private entity such as MNN who opens its property for speech by others is not transformed by that fact alone into a state actor. Under the text of the Constitution and our precedents, MNN is not a state actor subject to the First Amendment. We reverse in relevant part the judgment of the Second Circuit, and we remand the case for further proceedings consistent with this opinion.

It is so ordered.
JUSTICE SOTOMAYOR, with whom JUSTICE GINSBURG, JUSTICE BREYER, and JUSTICE KAGAN join, dissenting.

The Court tells a very reasonable story about a case that is not before us. I write to address the one that is.

This is a case about an organization appointed by the government to administer a constitutional public forum. (It is not, as the Court suggests, about a private property owner that simply opened up its property to others.) New York City (the City) secured a property interest in public-access television channels when it granted a cable franchise to a cable company. State regulations require those public-access channels to be made open to the public on terms that render them a public forum. The City contracted out the administration of that forum to a private organization, petitioner Manhattan Community Access Corporation (MNN). By accepting that agency relationship, MNN stepped into the City’s shoes and thus qualifies as a state actor, subject to the First Amendment like any other.

I

A cable-television franchise is, essentially, a license to create a system for distributing cable TV in a certain area. It is a valuable right, usually conferred on a private com-

Cable companies transmit content through wires that stretch “between a transmission facility and the television sets of individual subscribers.” *Id.*, at 627–628. Creating this network of wires is a disruptive undertaking that “entails the use of public rights-of-way and easements.” *Id.*, at 628.

New York State authorizes municipalities to grant cable franchises to cable companies of a certain size only if those companies agree to set aside at least one public access channel. 16 N. Y. Codes, Rules & Regs. §§895.1(f), 895.4(b)(1) (2016). New York then requires that those public-access channels be open to all comers on “a first-come, first-served, nondiscriminatory basis.” §895.4(c)(4). Likewise, the State prohibits both cable franchisees and local governments from “exercis[ing] any editorial control” over the channels, aside from regulating obscenity and other unprotected content. §§895.4(c)(8)–(9).

**B**

Years ago, New York City (no longer a party to this suit) and Time Warner Entertainment Company (never a party to this suit) entered into a cable-franchise agreement. App. 22. Time Warner received a cable franchise; the City received public-access channels. The agreement also provided that the public-access channels would be operated by an independent, nonprofit corporation chosen by the Manhattan borough president. But the City, as the practice of other New York municipalities confirms, could have instead chosen to run the channels itself. See §895.4(c)(1); Brief for Respondents 35 (citing examples).

MNN is the independent nonprofit that the borough president appointed to run the channels; indeed, MNN
appears to have been incorporated in 1991 for that precise purpose, with seven initial board members selected by the borough president (though only two thus selected today). See App. 23; Brief for Respondents 7, n. 1. The City arranged for MNN to receive startup capital from Time Warner and to be funded through franchise fees from Time Warner and other Manhattan cable franchisees. App. 23; Brief for New York County Lawyers Association (NYCLA) as Amicus Curiae 27; see also App. to Brief for Respondents 19a. As the borough president announced upon MNN’s formation in 1991, MNN’s “central charge is to administer and manage all the public access channels of the cable television systems in Manhattan.” App. to Brief for NYCLA as Amicus Curiae 1.

As relevant here, respondents DeeDee Halleck and Jesus Papoleto Melendez sued MNN in U. S. District Court for the Southern District of New York under 42 U. S. C. §1983. They alleged that the public-access channels, “[r]equired by state regulation and [the] local franchise agreements,” are “a designated public forum of unlimited character”; that the City had “delegated control of that public forum to MNN”; and that MNN had, in turn, engaged in viewpoint discrimination in violation of respondents’ First Amendment rights. App. 39.

The District Court dismissed respondents’ First Amendment claim against MNN. The U. S. Court of Appeals for the Second Circuit reversed that dismissal, concluding that the public-access channels “are public forums and that [MNN’s] employees were sufficiently alleged to be state actors taking action barred by the First Amendment.” 882 F. 3d 300, 301–302 (2018). Because the case before us arises from a motion to dismiss, respondents’ factual allegations must be accepted as true. Hernandez v. Mesa, 582 U. S. ___, ___ (2017) (per curiam) (slip op., at 1).
I would affirm the judgment below. The channels are clearly a public forum: The City has a property interest in them, and New York regulations require that access to those channels be kept open to all. And because the City (1) had a duty to provide that public forum once it granted a cable franchise and (2) had a duty to abide by the First Amendment once it provided that forum, those obligations did not evaporate when the City delegated the administration of that forum to a private entity. Just as the City would have been subject to the First Amendment had it chosen to run the forum itself, MNN assumed the same responsibility when it accepted the delegation.

When a person alleges a violation of the right to free speech, courts generally must consider not only what was said but also in what context it was said.

On the one hand, there are “public forums,” or settings that the government has opened in some way for speech by the public (or some subset of it). The Court’s precedents subdivide this broader category into various subcategories, with the level of leeway for government regulation of speech varying accordingly. See Minnesota Voters Alliance v. Mansky, 585 U. S. __, ___ (2018) (slip op., at 7). Compare Frisby v. Schultz, 487 U. S. 474, 480 (1988) (streets and public parks, traditional public forums), with Southeastern Promotions, Ltd. v. Conrad, 420 U. S. 546, 555 (1975) (city-leased theater, designated public forum), with Christian Legal Soc. Chapter of Univ. of Cal., Hastings College of Law v. Martinez, 561 U. S. 661, 669, 679, and n. 12 (2010) (program for registered student organizations, limited public forum). But while many cases turn on which type of “forum” is implicated, the important point here is that viewpoint discrimination is impermissible in them all. See Good News Club v. Milford Central
SOTOMAYOR, J., dissenting


On the other hand, there are contexts that do not fall under the “forum” rubric. For one, there are contexts in which the government is simply engaging in its own speech and thus has freedom to select the views it prefers. See, e.g., Walker v. Texas Div., Sons of Confederate Veter-
ans, Inc., 576 U. S. ___, ___–___ (2015) (slip op., at 6–7) (specialty license plates); Pleasant Grove City v. Summum, 555 U. S. 460, 467–469, 481 (2009) (privately donated permanent monuments in a public park).1 In addition, there are purely private spaces, where the First Amend-
ment is (as relevant here) inapplicable. The First Amendment leaves a private store owner (or homeowner), for example, free to remove a customer (or dinner guest) for expressing unwanted views. See, e.g., Lloyd Corp. v. Tanner, 407 U. S. 551, 569–570 (1972). In these settings, there is no First Amendment right against viewpoint discrimination.

Here, respondents alleged viewpoint discrimination. App. 39. So a key question in this case concerns what the Manhattan public-access channels are: a public forum of some kind, in which a claim alleging viewpoint discrimination would be cognizable, or something else, such as government speech or purely private property, where picking favored viewpoints is appropriately commonplace.2 Neither MNN nor the majority suggests that this is an in-

1 That does not mean that no restrictions apply at all to the government’s expression in such spaces, but it does mean that the government can pick and choose among different views. See Walker, 576 U. S., at ___–___ (slip op., at 6, 17–18); Summum, 555 U. S., at 468.

2 The channels are not, of course, a physical place. Under the Court’s precedents, that makes no difference: Regardless of whether something “is a forum more in a metaphysical than in a spatial or geographic sense, . . . the same principles are applicable.” Rosenberger v. Rector and Visitors of Univ. of Va., 515 U. S. 819, 830 (1995) (treating “Student Activities Fund” as the forum at issue and citing cases in which a school’s mail system and a charity drive were the relevant forums).
SOTOMAYOR, J., dissenting

stance of government speech. This case thus turns first and foremost on whether the public-access channels are or are not purely private property.\(^3\)

This Court has not defined precisely what kind of governmental property interest (if any) is necessary for a public forum to exist. See *Cornelius v. NAACP Legal Defense & Ed. Fund, Inc.*, 473 U. S. 788, 801 (1985) ("a speaker must seek access to public property or to private property dedicated to public use"). But see *ante*, at 11, n. 3 (appearing to reject the phrase "private property dedicated to public use" as "passing dicta"). I assume for the sake of argument in this case that public-forum analysis is inappropriate where the government lacks a "significant property interest consistent with the communicative purpose of the forum." *Denver Area Ed. Telecommunications Consortium, Inc. v. FCC*, 518 U. S. 727, 829 (1996) (THOMAS, J., concurring in judgment in part and dissenting in part).

Such an interest is present here. As described above, New York State required the City to obtain public-access channels from Time Warner in exchange for awarding a cable franchise. See *supra*, at 2. The exclusive right to use these channels (and, as necessary, Time Warner’s infrastructure) qualifies as a property interest, akin at the very least to an easement.

The last time this Court considered a case centering on public-access channels, five Justices described an interest like the one here as similar to an easement. Although JUSTICE BREYER did not conclude that a public-access channel was indeed a public forum, he likened the cable

\(^3\)As discussed below, it is possible that some (or even many) public-access channels are government speech. The channels that MNN administers, however, are clearly better thought of as a public forum given the New York regulations mandating open and equal access. See *infra*, at 9–10, and n. 7.
company’s agreement to reserve such channels “to the reservation of a public easement, or a dedication of land for streets and parks, as part of a municipality’s approval of a subdivision of land.” Denver Area, 518 U. S., at 760–761 (joined by Stevens and Souter, JJ.). And Justice Kennedy observed not only that an easement would be an appropriate analogy, id., at 793–794 (opinion concurring in part, concurring in judgment in part, and dissenting in part, joined by GINSBURG, J.), but also that “[p]ublic access channels meet the definition of a public forum,” id., at 791, “even though they operate over property to which the cable operator holds title,” id., at 792; see also id., at 792–793 (noting that the entire cable system’s existence stems from the municipality’s decision to grant the franchise). What those five Justices suggested in 1996 remains true today.

“A common idiom describes property as a ‘bundle of sticks’—a collection of individual rights which, in certain combinations, constitute property.” United States v. Craft, 535 U. S. 274, 278 (2002). Rights to exclude and to use are two of the most crucial sticks in the bundle. See id., at 283. “State law determines . . . which sticks are in a person’s bundle,” id., at 278, and therefore defining property itself is a state-law exercise. As for whether there is a sufficient property interest to trigger First Amendment forum analysis, related precedents show that there is.

As noted above, there is no disputing that Time Warner owns the wires themselves. See Turner, 512 U. S., at 628. If the wires were a road, it would be easy to define the public’s right to walk on it as an easement. See, e.g., In re India Street, 29 N. Y. 2d 97, 100–103, 272 N. E 2d 518, 518–519, 296 N. Y. 2d 521, 521–522, 296 N. E 2d 519, 519–520, 296 N. Y. 2d 521, 521–522, 296 N. E 2d 519, 519–520, 296 N. Y. 2d 521, 521–522, 296 N. E 2d 519, 519–520.

4 The parties have not pointed this Court to any New York law definitively establishing the status of the channels. But even if there were uncertainty about the status of the channels under New York law, that would not be a reason to resolve the case against respondents (plaintiffs below) at the motion to dismiss stage. See infra, at 12, n. 9, 14.
Similarly, if the wires were a theater, there would be no question that a government’s long-term lease to use it would be sufficient for public-forum purposes. *Southeastern Promotions*, 420 U. S., at 547, 555. But some may find this case more complicated because the wires are not a road or a theater that one can physically occupy; they are a conduit for transmitting signals that appear as television channels. In other words, the question is how to understand the right to place content on those channels using those wires.

The right to convey expressive content using someone else’s physical infrastructure is not new. To give another low-tech example, imagine that one company owns a billboard and another rents space on that billboard. The renter can have a property interest in placing content on the billboard for the lease term even though it does not own the billboard itself. See, e.g., *Naegele Outdoor Advertising Co. of Minneapolis v. Lakeville*, 532 N. W. 2d 249, 253 (Minn. 1995); see also *Matter of XAR Corp. v. Di Donato*, 76 App. Div. 2d 972, 973, 429 N. Y. S. 2d 59, 60 (1980) (“Although invariably labeled ‘leases,’ agreements to erect advertising signs or to place signs on walls or fences are easements in gross”).

The same principle should operate in this higher tech realm. Just as if the channels were a billboard, the City obtained rights for exclusive use of the channels by the public for the foreseeable future; no one is free to take the channels away, short of a contract renegotiation. Cf. *Craft*, 535 U. S., at 283. The City also obtained the right to administer, or delegate the administration of, the channels. The channels are more intangible than a billboard, but no one believes that a right must be tangible to qualify as a property interest. See, e.g., *Armstrong v. United States*, 364 U. S. 40, 48–49 (1960) (treating destruction of valid liens as a taking); *Adams Express Co. v. Ohio State Auditor*, 166 U. S. 185, 219 (1897) (treating “privileges,
corporate franchises, contracts or obligations” as taxable property). And it is hardly unprecedented for a government to receive a right to transmit something over a private entity’s infrastructure in exchange for conferring something of value on that private entity; examples go back at least as far as the 1800s.5

I do not suggest that the government always obtains a property interest in public-access channels created by franchise agreements. But the arrangement here is consistent with what the Court would treat as a governmental property interest in other contexts. New York City gave Time Warner the right to lay wires and sell cable TV. In exchange, the City received an exclusive right to send its own signal over Time Warner’s infrastructure—no different than receiving a right to place ads on another’s billboards. Those rights amount to a governmental property interest in the channels, and that property interest is clearly “consistent with the communicative purpose of the forum,” Denver Area, 518 U. S., at 829 (opinion of THOMAS, J.). Indeed, it is the right to transmit the very content to which New York law grants the public open and equal access.

With the question of a governmental property interest resolved, it should become clear that the public-access channels are a public forum.6 Outside of classic examples

5For example, during the railroad boom, governments obtained not only physical easements in favor of the public over tracks used, owned, and managed by private railroads, including rights to use the rails and all relevant “fixtures and appurtenances,” see, e.g., Lake Superior & Mississippi R. Co. v. United States, 93 U. S. 442, 444, 453–454 (1877), but also, in some situations, rights to transmit personnel and freight for free or at reduced rates, Ellis, Railroad Land Grant Rates, 1850–1945, 21 J. Land & P. U. Econ. 207, 209, 211–212 (1945).

6Though the majority disagrees on the property question, I do not take it seriously to dispute that this point would follow. See ante; at
like sidewalks and parks, a public forum exists only where the government has deliberately opened up the setting for speech by at least a subset of the public. *Cornelius*, 473 U. S., at 802. “Accordingly, the Court has looked to the policy and practice of the government,” as well as the nature of the property itself, “to ascertain whether it intended to designate a place not traditionally open to assembly and debate as a public forum.” See *ibid.* For example, a state college might make its facilities open to student groups, or a municipality might open up an auditorium for certain public meetings. See *id.*, at 802–803.

The requisite governmental intent is manifest here. As noted above, New York State regulations require that the channels be made available to the public “on a first-come, first-served, nondiscriminatory basis.” 16 N. Y. Codes, Rules & Regs. §895.4(c)(4); see also §§895.4(c)(8)–(9). The State, in other words, mandates that the doors be wide open for public expression. MNN’s contract with Time Warner follows suit. App. 23. And that is essentially how MNN itself describes things. See Tr. of Oral Arg. 9 (“We do not prescreen videos. We—they come into the door. We put them on the air”). These regulations “evidenc[e] a clear intent to create a public forum.” *Cornelius*, 473 U. S., at 802.

B

If New York’s public-access channels are a public forum, it follows that New York cannot evade the First Amendment by contracting out administration of that forum to a

---

7 New York may be uncommon (as it often is); public-access channels in other States may well have different policies and practices that make them more like government speech than constitutional forums. See Brief for Respondents 30–31; Brief for American Civil Liberties Union et al. as Amici Curiae 13–15. New York’s scheme, however, is the only one before us.
private agent. When MNN took on the responsibility of administering the forum, it stood in the City’s shoes and became a state actor for purposes of 42 U. S. C. §1983.

This conclusion follows from the Court’s decision in *West v. Atkins*, 487 U. S. 42 (1988). The Court in *West* unanimously held that a doctor hired to provide medical care to state prisoners was a state actor for purposes of §1983. *Id.*, at 54; see also *id.*, at 58 (Scalia, J., concurring in part and concurring in judgment). Each State must provide medical care to prisoners, the Court explained, *id.*, at 54, and when a State hires a private doctor to do that job, the doctor becomes a state actor, “‘clothed with the authority of state law,’” *id.*, at 55. If a doctor hired by the State abuses his role, the harm is “caused, in the sense relevant for state-action inquiry,” by the State’s having incarcerated the prisoner and put his medical care in that doctor’s hands. *Ibid.*

The fact that the doctor was a private contractor, the Court emphasized, made no difference. *Ibid.* It was “the physician’s function within the state system,” not his private-contractor status, that determined whether his conduct could “fairly be attributed to the State.” *Id.*, at 55–56. Once the State imprisoned the plaintiff, it owed him duties under the Eighth Amendment; once the State delegated those duties to a private doctor, the doctor became a state actor. See *ibid.*; see also *id.*, at 56–57. If the rule were any different, a State would “be free to contract out all services which it is constitutionally obligated to provide and leave its citizens with no means for vindication of those rights, whose protection has been delegated to ‘private’ actors, when they have been denied.” *Id.*, at 56, n. 14.

*West* resolves this case. Although the settings are different, the legal features are the same: When a government (1) makes a choice that triggers constitutional obligations, and then (2) contracts out those constitutional
responsibilities to a private entity, that entity—in agreeing to take on the job—becomes a state actor for purposes of §1983.8

Not all acts of governmental delegation necessarily trigger constitutional obligations, but this one did. New York State regulations required the City to secure public-access channels if it awarded a cable franchise. 16 N. Y. Codes, Rules & Regs. §895.4(b)(1). The City did award a cable franchise. The State’s regulations then required the City to make the channels it obtained available on a “first-come, first-served, nondiscriminatory basis.”9 §895.4(c)(4).

8 Governments are, of course, not constitutionally required to open prisons or public forums, but once they do either of these things, constitutional obligations attach. The rule that a government may not evade the Constitution by substituting a private administrator, meanwhile, is not a prison-specific rule. More than 50 years ago, for example, this Court made clear in Evans v. Newton, 382 U. S. 296 (1966), that the city of Macon, Georgia, could not evade the Fourteenth Amendment’s Equal Protection Clause by handing off control of a park to a group “of ‘private’ trustees.” Id., at 301. Rather, “the public character of [the] park require[d] that it be treated as a public institution subject to the command of the Fourteenth Amendment, regardless of who ha[d] title under state law.” Id., at 302.

9 Accordingly, this is not a case in which a private entity has been asked to exercise standardless discretion. See, e.g., American Mfrs. Mut. Ins. Co. v. Sullivan, 526 U. S. 40, 52 (1999). Had New York law left MNN free to choose its favorite submissions, for example, a different result might well follow.

MNN has suggested to this Court that its contract with Time Warner allows it “to curate content, to decide to put shows together on one of our channels or a different channel.” Tr. of Oral Arg. 6; see Reply Brief 9. But MNN’s contract cannot defeat New York law’s “first-come, first-served, nondiscriminatory” scheduling requirement, 16 N. Y. Codes, Rules & Regs. §895.4(c)(4), and the discretion MNN asserts seems to be at most some limited authority to coordinate the exact placement and timing of the content it is obliged to accept indiscriminately, see Tr. of Oral Arg. 25–26. That seems akin to the authority to make reasonable time, place, and manner provisions, which is consistent with administering any public forum. See Ward v. Rock Against Racism, 491 U. S. 781, 791 (1989). As for any factual assertions about how the channels
That made the channels a public forum. See *supra*, at 9–10. Opening a public forum, in turn, entailed First Amendment obligations.

The City could have done the job itself, but it instead delegated that job to a private entity, MNN. MNN could have said no, but it said yes. (Indeed, it appears to exist entirely to do this job.) By accepting the job, MNN accepted the City’s responsibilities. See *West*, 487 U. S., at 55. The First Amendment does not fall silent simply because a government hands off the administration of its constitutional duties to a private actor.

III

The majority acknowledges that the First Amendment could apply when a local government either (1) has a property interest in public-access channels or (2) is more directly involved in administration of those channels than the City is here. *Ante*, at 15. And it emphasizes that it “decide[s] only the case before us in light of the record before us.” *Ibid*. These case-specific qualifiers sharply limit the immediate effect of the majority’s decision, but that decision is still meaningfully wrong in two ways. First, the majority erroneously decides the property question against the plaintiffs as a matter of law. Second, and more fundamentally, the majority mistakes a case about the government choosing to hand off responsibility to an agent for a case about a private entity that simply enters a marketplace.

A

The majority’s explanation for why there is no govern-
mental property interest here, ante, at 14–15, does not hold up. The majority focuses on the fact that “[b]oth Time Warner and MNN are private entities”; that Time Warner “owns its cable network, which contains the public access channels”; and that “MNN operates those public access channels with its own facilities and equipment.” Ante, at 14; see also ante, at 15. Those considerations cannot resolve this case. The issue is not who owns the cable network or that MNN uses its own property to operate the channels. The key question, rather, is whether the channels themselves are purely private property. An advertiser may not own a billboard, but that does not mean that its long-term lease is not a property interest. See supra, at 8.

The majority also says that “[n]othing in the record here suggests that a government . . . owns or leases either the cable system or the public access channels at issue here.” Ante, at 14. But the cable system itself is irrelevant, and, as explained above, the details of the exchange that yielded Time Warner’s cable franchise suggest a governmental property interest in the channels. See supra, at 6–9.

The majority observes that “the franchise agreements expressly place the public access channels ‘under the jurisdiction’ of MNN,” ante, at 14, but that language simply describes the City’s appointment of MNN to administer the channels. The majority also chides respondents for failing to “alleg[e] in their complaint that the City has a property interest in the channels,” ibid., but, fairly read, respondents’ complaint includes such an assertion.10 In

10Respondents alleged that the City “created an electronic public forum” and “delegat[ed] control of that forum to” MNN. App. 17. They further alleged that “[a]ll cable franchise agreements require cable operators—as a condition for easements to use the public rights-of-way—to dedicate some channels for programming by the public,” id., at 20, invoked the state regulations requiring the designation of a channel here, id., at 21, and then alleged that the City’s franchise
any event, any ambiguity or imprecision does not justify resolving the case against respondents at the motion-to-dismiss stage. To the extent the majority has doubts about respondents’ complaint—or factual or state-law issues that may bear upon the existence of a property interest—the more prudent course would be to vacate and remand for the lower courts to consider those matters more fully. In any event, as I have explained, the best course of all would be to affirm.

B

More fundamentally, the majority’s opinion erroneously fixates on a type of case that is not before us: one in which a private entity simply enters the marketplace and is then subject to government regulation. The majority swings hard at the wrong pitch.

The majority focuses on *Jackson v. Metropolitan Edison Co.*, 419 U. S. 345 (1974), which is a paradigmatic example of a line of cases that reject §1983 liability for private actors that simply operate against a regulatory backdrop. *Jackson* emphasized that the “fact that a business is subject to state regulation does not by itself convert its action into that of the State.” *Id.*, at 350; accord, *ante*, at 12. Thus, the fact that a utility company entered the marketplace did not make it a state actor, even if it was highly regulated. See *Jackson*, 419 U. S., at 358; accord, agreement “requires Time Warner to set aside” the channels, *id.*, at 22. While the complaint does not use the words “property interest,” those allegations can be read to include the idea that whatever was “set aside” or “dedicate[d],” *id.*, at 20, 22, qualified as a sufficient City property interest to support respondents’ assertion of a public forum. Cf. *People v. Brooklyn & Queens Transit Corp.*, 273 N. Y. 394, 400–401, 7 N. E. 2d 833, 835 (1937) (discussing dedications of property to public use); cf. also *Denver Area Ed. Telecommunications Consortium, Inc. v. FCC*, 518 U. S. 727, 794 (1996) (Kennedy, J., concurring in part, concurring in judgment in part, and dissenting in part) (noting this theory).
The same rule holds, of course, for private comedy clubs and grocery stores. See ante, at 9.11

The Jackson line of cases is inapposite here. MNN is not a private entity that simply ventured into the marketplace. It occupies its role because it was asked to do so by the City, which secured the public-access channels in exchange for giving up public rights of way, opened those channels up (as required by the State) as a public forum, and then deputized MNN to administer them. That distinguishes MNN from a private entity that simply sets up shop against a regulatory backdrop. To say that MNN is nothing more than a private organization regulated by the government is like saying that a waiter at a restaurant is

11 There was a time when this Court’s precedents may have portended the kind of First Amendment liability for purely private property owners that the majority spends so much time rejecting. See Marsh v. Alabama, 326 U. S. 501, 505–509 (1946) (treating a company-owned town as subject to the First Amendment); Food Employees v. Logan Valley Plaza, Inc., 391 U. S. 308, 315–320, and n. 9, 325 (1968) (extending Marsh to cover a private shopping center to the extent that it sought to restrict speech about its businesses). But the Court soon stanched that trend. See Lloyd Corp. v. Tanner, 407 U. S. 551, 561–567 (1972) (cabining Marsh and refusing to extend Logan Valley); Hudgens v. NLRB, 424 U. S. 507, 518 (1976) (making clear that “the rationale of Logan Valley did not survive” Lloyd). Ever since, this Court has been reluctant to find a “public function” when it comes to “private commercial transactions” (even if they occur against a legal or regulatory backdrop), see, e.g., Flagg Bros., Inc. v. Brooks, 436 U. S. 149, 161–163 (1978), instead requiring a closer connection between the private entity and a government or its agents, see, e.g., Brentwood Academy v. Tennessee Secondary School Athletic Assn., 531 U. S. 288, 298 (2001) (nonprofit interscholastic athletic association “pervasive[ly] entwine[d]” with governmental institutions and officials); Lugar v. Edmondson Oil Co., 457 U. S. 922, 942 (1982) (state-created system “whereby state officials [would] attach property on the ex parte application of one party to a private dispute”); see also Burton v. Wilmington Parking Authority, 365 U. S. 715, 723–725 (1961) (restaurant in municipal parking garage partly maintained by municipal agency); accord, ante, at 6–7. Jackson exemplifies the line of cases that supplanted cases like Logan Valley—not cases like this one.
an independent food seller who just happens to be highly regulated by the restaurant’s owners.

The majority also relies on the Court’s statements that its “public function” test requires that a function have been “traditionally and exclusively performed” by the government. *Ante*, at 6 (emphasis deleted); see *Jackson*, 419 U. S., at 352. Properly understood, that rule cabins liability in cases, such as *Jackson*, in which a private actor ventures of its own accord into territory shared (or regulated) by the government (e.g., by opening a power company or a shopping center). The Court made clear in *West* that the rule did not reach further, explaining that “the fact that a state employee’s role parallels one in the private sector” does not preclude a finding of state action. 487 U. S., at 56, n. 15.

When the government hires an agent, in other words, the question is not whether it hired the agent to do something that can be done in the private marketplace too. If that were the key question, the doctor in *West* would not have been a state actor. Nobody thinks that orthopedics is a function “traditionally exclusively reserved to the State,” *Jackson*, 419 U. S., at 352.

The majority consigns *West* to a footnote, asserting that its “scenario is not present here because the government has no [constitutional] obligation to operate public access channels.” *Ante*, at 7, n. 1. The majority suggests that *West* is different because “the State was constitutionally obligated to provide medical care to prison inmates.” *Ante*, at 7, n. 1. But what the majority ignores is that the State in *West* had no constitutional obligation to open the prison or incarcerate the prisoner in the first place; the obligation to provide medical care arose when it made those prior choices.

The City had a comparable constitutional obligation here—one brought about by its own choices, made against a state-law backdrop. The City, of course, had no constitut-
tional obligation to award a cable franchise or to operate public-access channels. But once the City did award a cable franchise, New York law required the City to obtain public-access channels, see *supra*, at 2, and to open them up as a public forum, see *supra*, at 9–10. That is when the City’s obligation to act in accordance with the First Amendment with respect to the channels arose. That is why, when the City handed the administration of that forum off to an agent, the Constitution followed. See *supra*, at 10–13.\(^\text{12}\)

The majority is surely correct that “when a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment.” *Ante*, at 9. That is because the majority is not talking about *constitutional* forums—it is talking about spaces where private entities have simply invited others to come speak. A comedy club can decide to open its doors as wide as it wants, but it cannot appoint itself as a government agent. The difference is between providing a service of one’s own accord and being asked by the government to administer a constitutional responsibility (indeed, here, existing to do so) on the government’s behalf.\(^\text{13}\)

\(^\text{12}\) *Jackson v. Metropolitan Edison Co.*., 419 U. S. 345 (1974), by contrast, exemplifies a type of case in which a private actor provides a service that there is no governmental obligation to provide at all. See *id.*, at 353 (no state requirement for government to provide utility service); see also, *e.g.*, *Hudgens*, 424 U. S. 507 (shopping center). In *West v. Atkins*, 487 U. S. 42 (1988), by contrast, the prison was obligated to provide health care in accordance with the Eighth Amendment to its prisoners once it incarcerated them, and here, the City was required to provide a public forum to its residents in accordance with the First Amendment once it granted the cable franchise. See *supra*, at 11–13.

\(^\text{13}\) Accordingly, the majority need not fear that “all private property owners and private lessees who open their property for speech [c]ould be subject to First Amendment constraints.” *Ante*, at 10. Those kinds of entities are not the government’s agents; MNN is. Whether such entities face “extensive regulation” or require “government licenses, government contracts, or government-granted monopolies,” *ante*, at 12,
To see more clearly the difference between the cases on which the majority fixates and the present case, leave aside the majority’s private comedy club. Imagine instead that a state college runs a comedy showcase each year, renting out a local theater and, pursuant to state regulations mandating open access to certain kinds of student activities, allowing students to sign up to perform on a first-come, first-served basis. Cf. *Rosenberger v. Rector and Visitors of Univ. of Va.*, 515 U. S. 819 (1995). After a few years, the college decides that it is tired of running the show, so it hires a performing-arts nonprofit to do the job. The nonprofit prefers humor that makes fun of a certain political party, so it allows only student acts that share its views to participate. Does the majority believe that the nonprofit is indistinguishable, for purposes of state action, from a private comedy club opened by local entrepreneurs?

I hope not. But two dangers lurk here regardless. On the one hand, if the City’s decision to outsource the channels to a private entity did render the First Amendment irrelevant, there would be substantial cause to worry about the potential abuses that could follow. Can a state university evade the First Amendment by hiring a nonprofit to apportion funding to student groups? Can a city do the same by appointing a corporation to run a municipal theater? What about its parks?

On the other hand, the majority hastens to qualify its decision, see *ante*, at 7, n. 1, 15, and to cabin it to the specific facts of this case, *ante*, at 15. Those are prudent limitations. Even so, the majority’s focus on *Jackson* still risks sowing confusion among the lower courts about how and when government outsourcing will render any abuses that follow beyond the reach of the Constitution.

In any event, there should be no confusion here. MNN

———
is immaterial, so long as they have not accepted the government’s request to fulfill the government’s duties on its behalf.
SOTOMAYOR, J., dissenting

is not a private entity that ventured into the marketplace and found itself subject to government regulation. It was asked to do a job by the government and compensated accordingly. If it does not want to do that job anymore, it can stop (subject, like any other entity, to its contractual obligations). But as long as MNN continues to wield the power it was given by the government, it stands in the government’s shoes and must abide by the First Amendment like any other government actor.

IV

This is not a case about bigger governments and smaller individuals, ante, at 16; it is a case about principals and agents. New York City opened up a public forum on public-access channels in which it has a property interest. It asked MNN to run that public forum, and MNN accepted the job. That makes MNN subject to the First Amendment, just as if the City had decided to run the public forum itself.

While the majority emphasizes that its decision is narrow and factbound, ante, at 15, that does not make it any less misguided. It is crucial that the Court does not continue to ignore the reality, fully recognized by our predecessors, that private actors who have been delegated constitutional responsibilities like this one should be accountable to the Constitution’s demands. I respectfully dissent.
U.N. Special Rapporteur’s Latest Report on Online Content Regulation Calls for ‘Human Rights by Default’

By Evelyn Douek   Wednesday, June , 2018, 8:00 AM

David ÿe, the U.N. special rapporteur on the promotion and protection of the right to freedom of opinion and expression, released his latest report to the U.N. Human Rights Council last week. The report calls for states and companies to apply international human rights law at all stages of online content regulation: from creating rules about what content should be taken down, to conducting due diligence about how changes to platforms affect human rights, to providing remedies for people harmed by moderation decisions.

This is the first U.N. report to examine the regulation of user-generated online content. It could hardly be more timely. The ongoing reckoning over social media has been one of the most pervasive stories of 2018 as nation-states and companies grapple with social media’s effects on communities and democratic institutions both online and offline.

So far, however, the changes prompted by this reckoning have been patchwork. Individual social media companies are implementing voluntary initiatives to try to quell concerns and fend off further regulation. In the past few months alone, Facebook has created transparency measures for political advertising, publicized the internal guidelines it uses to make decisions under its community standards about what content stays up and comes down, announced a new appeals process, and, for the first time, published numbers relating to the enforcement of those community standards. Twitter has also announced ways it intends to regulate political advertising and efforts to help foster better “conversational health” on its platform. These are merely prominent examples of a general trend.

However well-intentioned those actions are, these companies are trying to use internal improvements to get ahead of further state regulation. So far, they are seeing varying degrees of success in persuading regulators that new laws are unnecessary. As I wrote last month, the European Union continues to threaten social media platforms with heavier regulation. A U.K parliamentary inquiry into fake news rumbles on and continues to be aggressive in charging platforms with lax practices.

These are just two examples against the background of a global increase in governmentally imposed obligations to monitor and remove user-generated content. ÿe’s report cites recent Chinese, German, EU and ÿiyian examples in two short paragraphs alone. There is little coordination or collaboration in the rollout of these public and private measures as companies and regulators alike focus on their individual concerns rather than the global online ecosystem. ÿe’s report, therefore, offers a rare international lens on the intractable problem of content regulation. This is much-needed insight in the nascent, interconnected world of online public spheres. The framework the report proposes, based on international human rights law, offers a way of ensuring greater consistency and transparency in the different environments that social media companies operate.

For states, respecting human rights law means their laws should not unduly restrict freedom of expression, either online or offline. Many of ÿe’s previous reports have focused on these obligations. As his latest report argues, this includes refraining from imposing disproportionate liability on social media companies because this creates an incentive for companies to over-censor content, which has a chilling effect on freedom of expression. Such laws also delegate responsibility for censorship decisions to private companies, rather than using public legal processes that comply with the rule of law.

For most companies, the report’s recommendation that they adopt international human rights law as the authoritative standard for their content moderation would be a significant change in their operating model.

International Norms: Human Rights by Default'

As ÿe’s report explains, most companies do not explicitly base their content standards on any single body of law. They retain large amounts of discretion and generally regulate content according to their own terms of service, subject to complying with local laws in the jurisdictions where they operate. This commitment to local legal compliance can cause problems where the local laws are vague, where they are themselves inconsistent with human rights law or where they are not sufficient to protect human rights. In these circumstances, company decisions are often driven by commercial considerations and the extent to which local governments can apply effective pressure. Where a state is strong, or a market is valuable, this can make users vulnerable to violations of their rights when a country insists on censorship that does not accord with human rights law.

Irrespective, ÿe identifies Germany’s new “NetzDG” law, which imposes extremely high potential penalties and requires tight time frames for removal, as raising this concern. Other violations may occur where a strong state seeks content removals outside legal processes or using arrangements that have limited transparency. ÿe points to an instance in which Pakistan compelled Google to offer a local
version of YouTube that removed content the government found offensive and cited an agreement between Facebook and Israel to remove content the government flagged as “incitement.”

On the other hand, in places where a state is weak or business in a given country is not particularly valuable to the company, there may be insufficient pressure or commercial incentive for companies to monitor their platforms and preserve healthy speech environments. The most notorious example of this is Myanmar, where hate speech has fueled ethnic violence and genocide while Facebook devoted fairly limited resources to content moderation. In Sri Lanka, as the New York Times reported, the government struggled to get Facebook’s attention regarding viral hate speech in the context of escalating ethnic violence—until the government blocked most social media in the country.

The result is that these transnational companies apply an opaque and variable kind of regulation that Kye calls “platform law,” which allows platforms a great deal of discretion with little accountability. Human rights law, he argues, offers a solution:

Private norms, which vary according to each company’s business model and vague assertions of community interests, have created unstable, unpredictable and unsafe environments for users and intensified government scrutiny. National laws are inappropriate for companies that seek common norms for their geographically and culturally diverse user base. But human rights standards, if implemented transparently and consistently with meaningful user and civil society input, provide a framework for holding both States and companies accountable to users across national borders.

Kye calls for companies to move from individual “platform law” to “human rights by default.”

As I have written before, there is little to compel transnational companies to observe international law. As non-state actors, they are not parties to international human rights treaties. The U.N. Guiding Principles on Business and Human Rights were developed to provide a framework for holding multinational corporations accountable for their impact on human rights, but they are not binding. Kye argues that the companies “overwhelming role in public life globally” is a strong reason for adopting the guiding principles. But whatever the normative appeal of these guidelines, they do not have legal force.

The report offers another reason, however, that the companies should adopt these norms. Abiding by international human rights norms “enables forceful normative responses against undue State restrictions.” This means that if countries pressure companies to enforce content standards in a way that is idiosyncratic or infringes users’ freedom of expression, companies are on stronger ground if they can appeal to international norms as their reason for resisting these demands.

Facebook’s founder and CEO, Mark Zuckerberg, has seemingly expressed a desire to find just such a higher norm. He has voiced discomfort with being put in the position of deciding where the line is on hate speech in different contexts and has even suggested a kind of Facebook Supreme Court to make decisions in hard cases. Kye’s report is a reminder that each company doesn’t necessarily need to create a new legal framework from whole cloth—one already exists in the form of international human rights law.

Kye’s report explains other benefits of adopting global standards beyond the substantive protection of human rights. For users, companies being guided by the same set of norms regarding freedom of expression in every country would offer greater consistency and predictability across markets and platforms. Furthermore, human rights law not only protects freedom of expression itself but also provides standards for transparency, due diligence about companies’ human rights impacts and due process for remediation of harms.

**Objections to Human Rights Law**

But international human rights norms are not a panacea for the intractable problems of online speech regulation. It is something of a misnomer to speak of international human rights law as if it is a single, self-contained and cohesive body of rules. Instead, these laws are found in a variety of international and regional treaties that are subject to differing interpretations by states that are parties to the conventions as well as international tribunals applying the laws.

In a recent comprehensive survey of international laws regulating speech, for example, Amal Clooney and Philippa Webb showed that different sources provide conflicting guidance on speech rights and have been interpreted inconsistently by different tribunals. These norms may not supply as much certainty and uniformity as promised or expected. Even if there were no problems of vagueness or conflicting authorities, international law on freedom of expression is not universally venerated: Clooney and Webb argue that the right to insult is not sufficiently protected and that while technology companies “should be able to draw inspiration from international law,” that law needs legal reform to play a more positive role.
More generally, the particular U.S. objections to international norms on freedom of expression are well known. The absolutism of the First Amendment is an outlier even in the democratic world, and the United States has entered reservations to international treaties containing speech rights—such as the International Covenant on Civil and Political Rights—to make clear that it will not be bound by international laws more restrictive of speech than the First Amendment. As Knicke's research has shown, content moderation policies at the dominant U.S. social media platforms were developed by lawyers acculturated in this First Amendment ethos.

But these platforms already depart from free-speech absolutism, removing content that the government could not proscribe under the First Amendment, for example harassment and nudity. Knicke also shows that in the years since their founding, the platforms gradually developed more nuanced approaches to regulation in countries with different speech standards than those in the U.S. The application of more universal human rights norms would not defeat this need to respect local context in determining how best to regulate speech. As Kye writes, human rights principles can offer a predictable and consistent baseline but are not "so inflexible or dogmatic" as to deprive companies of the ability to consider relevant context. Importantly, though, they would provide a "common vocabulary" for explaining decisions, which allows for greater dialogue and the development of a kind of precedent to help move toward better and more uniform decision making.

**Thinking Long Term and Big Picture**

As Kye's report and ongoing media coverage make clear, there is a growing trend of increased regulation of social media globally. States should adhere to their international legal obligations to respect individuals' rights to freedom of expression, regardless of frontiers. Where they do not, however, companies will need tools to allow their platforms to enable the exercise of human rights rather than instead enabling the violation of users' rights by governments. Transparent commitment to human rights as guiding content moderation would certainly be a voluntary constraint on their discretion—but it would also provide normative backing for their decisions.

No legal system is perfect. International human rights law, in particular, is dynamic and still in development. But just as the vaguely worded First Amendment has crystallized into more concrete rules, so too can international law. Compared with the First Amendment, international law on freedom of expression is young, having come into existence only since World War II, and active engagement with these norms and transparency around decisions could facilitate its ongoing development. Kye suggests that companies can develop a kind of case law by explaining their decisions, which would enable users, civil society and states to better understand their standards. He further advocates an industry-wide social media council, modeled on press councils, to provide an accountability mechanism and to be a credible and independent body to help develop transparency and norms.

"Content regulation" is an umbrella term that encompasses a wide range of complex issues, which play out on different platforms run by a diverse range of companies in countries around the world. Each of these issues requires sustained individual attention, but it's also important not to miss the forest for the trees. Kye's report offers a framework based on process and transparency that takes as its starting point this bigger picture—and, within it, the intrinsic value of a greater harmonization of norms across companies and countries.
What is platform governance?

Robert Gorwa

To cite this article: Robert Gorwa (2019) What is platform governance?, Information, Communication & Society, 22:6, 854-871, DOI: 10.1080/1369118X.2019.1573914

To link to this article: https://doi.org/10.1080/1369118X.2019.1573914

Published online: 11 Feb 2019.

Submit your article to this journal

Article views: 1689

View Crossmark data
What is platform governance?

Robert Gorwa

Department of Politics and International Relations, University of Oxford, Oxford, UK

ABSTRACT
Following a host of high-profile scandals, the political influence of platform companies (the global corporations that operate online ‘platforms’ such as Facebook, WhatsApp, YouTube, and many other online services) is slowly being re-evaluated. Amidst growing calls to regulate these companies and make them more democratically accountable, and a host of policy interventions that are actively being pursued in Europe and beyond, a better understanding of how platform practices, policies, and affordances (in effect, how platforms govern) interact with the external political forces trying to shape those practices and policies is needed. Building on digital media and communication scholarship as well as governance literature from political science and international relations, the aim of this article is to map an interdisciplinary research agenda for platform governance, a concept intended to capture the layers of governance relationships structuring interactions between key parties in today’s platform society, including platform companies, users, advertisers, governments, and other political actors.

ARTICLE HISTORY
Received 31 October 2018
Accepted 17 January 2019

KEYWORDS
Platform governance; platform; regulation; Facebook; Google; accountability

Introduction

Platform companies, the global corporations that operate Facebook, WhatsApp, YouTube, and many other online services, have become enmeshed in virtually all aspects of contemporary life, from politics (Gillespie, 2018b) and labour relations (Srnicek, 2016; Van Doorn, 2017) to cultural production and consumption (Nieborg & Poell, 2018). Although the products and services provided by a handful of large, predominantly American technology firms were not long ago widely portrayed as inherently beneficial and democratizing ‘liberation technologies’ (Tucker, Theocharis, Roberts, & Barberá, 2017), the influence of these corporations in public and political life is slowly being re-evaluated following multiple high-profile public scandals (Vaidhyanathan, 2018). Today, scholars, policymakers, and the public are increasingly attempting to understand the complex political effects of Instagram, Google Search, Airbnb, Uber, Amazon, and other platforms, amidst calls to make the companies that operate them more democratically accountable (Suzor, 2019).

This endeavour is the latest chapter in the effort to better understand the multitude of complex socio-technical developments playing a role in modern life. Research documenting the growing impact of automated decision-making in areas such as policing, finance, and healthcare (Ananny & Crawford, 2018) has been coupled with a constant flow of
‘algorithmic war stories’ illustrating how bias and discrimination can be exhibited by these systems (Edwards & Veale, 2017, p. 14). A growing and increasingly well-organized interdisciplinary community of social and computer scientists is responding by proposing actionable frameworks for accountable algorithms and fair machine learning (Barocas, Hardt, & Narayanan, 2018; Wallach, 2014), and how various desirable principles, such as ‘explainability’ or broader notions of legal ‘recourse,’ could be enacted for algorithmic systems (Ustun, Spangher, & Liu, 2019; Wachter, Mittelstadt, & Russell, 2017).

A similarly focused line of research into the governance of platform companies has yet to be pursued, despite the even wider prevalence of what might be termed ‘platform war stories.’ From the Cambridge Analytica scandal and Russian interference in the 2016 US election to Facebook’s troubling role in the recent Myanmar conflict, policymakers around the world are grappling with what Pasquale (2016, p. 314) has called ‘the darker narrative of platform capitalism.’ Popular discourse has become increasingly suffused with various regulatory proposals for curbing the power of the so-called ‘digital giants,’ which range from calls to break up Facebook from its acquisitions Instagram and WhatsApp (Reich, 2018) to arguments that platforms should be legally responsible for content posted by users via their services (Syal, 2017). How should platforms be governed? There is no simple answer, because, as Helberger, Pierson, and Poell (2018) outline, the policy arena is fragmented, with responsibility for the social and political role of platforms divided between the platform companies (as architects of online environments), users (as individuals making decisions about their specific behaviour in an online environment), and governments (as the entities setting the overall ground rules for those interactions).

Although key questions about the appropriate balance of responsibilities between these actors remain unanswered, policymakers – at least in Europe – are keen to act, and Germany’s 2017 ‘Network Enforcement’ law (NetzDG), which makes companies liable for illegal speech propagated via their services, has provided important precedent (and to some, a warning sign) for possible government responses (Schulz, 2018). At stake are major issues of freedom of expression, political participation, and democratic governance, with tremendous implications for the future of digital communication. Yet we lack a shared vocabulary or framework for approaching the oncoming challenges.

The goal of this article is to bring governance scholarship in conversation with the emerging platform studies literature, and in doing so, contribute a more precise understanding of what is now being informally called ‘platform governance.’ It proposes a framework explicitly mindful of power dynamics and the effects of external political forces on the platform ecosystem and seeks to provide a first mapping of the ongoing, interdisciplinary research landscape in a politically vital area that has yet to receive a concrete definition or theoretical overview. In the first section, I outline platform governance as a concept that refers to the layers of governance relationships structuring interactions between key parties in today’s platform society, capturing the growing body of work addressing the political effects of digital platforms (governance by platforms), as well as the complex challenges that the governance of platform companies presents. In the second section, three governance ‘lenses’ currently being used to inform thinking about platform regulation and policy are outlined, with a discussion of the questions and challenges raised by each. The final section addresses the normative question of how platforms should be governed going forward, and briefly discusses the emergent guiding principles shaping the future platform policy landscape.
Theorizing platform governance

What is a platform?

‘Platform’ is an ambiguous term, used differently by various scholarly communities (Schwarz, 2017): there are subtle variations in the ways that platforms are understood by computer scientists, economists, digital media scholars, and lawyers. A full history could go back to the emergence of computer networking, ARPANET, and early bulletin boards (Hauben & Hauben, 1997; Naughton, 2000), or even further back to the early emergence of computing (Hicks, 2017), but an abbreviated history of platforms might begin in California in the 1990s, as software developers began to conceptualize their offerings as more than just narrow programmes, but rather as flexible platforms that enable code to be developed and deployed (Bogost & Montfort, 2007). The term was then strategically deployed by certain companies, allowing them to brand themselves as platforms that facilitate access to user-generated content, but do not create it, and therefore should not be held liable for it (Gillespie, 2010). In recent years, the term has been adopted as shorthand both for the services provided by many technology companies, as well as the companies themselves (Srnicek, 2016). In this article, I use platform company more generally to refer to the corporations that deploy a service (e.g., Facebook and Alphabet) and platform to refer to those online, data-driven apps and services (e.g., Facebook Messenger, Google Search, YouTube); see Helmond (2015) for a more thorough discussion of history and definitions.

What is platform governance?

The political concept of governance has evolved greatly in the past half-century. Initially associated with domestic governments, governance was less a set of practices than a capacity: as per Fukuyama’s (2013, p. 4) traditional articulation, governance is the ‘government’s ability to make and enforce rules, and to deliver services.’ So-called ‘good governance,’ as commonly understood by political scientists, referred to a state’s ability to build functional and effective institutions, and use those institutions to maintain law and order (Weiss, 2000). However, a movement in the 1990s towards ‘global governance’ undertaken by a group of international relations scholars and political scientists, as famously catalyzed by a landmark 1992 volume, Governance without Government (Rosenau & Czempiel, 1992), advocated a much broader understanding of governance. This more flexible conception engaged with the central question of ‘how global life is organized, structured, and regulated’ (Barnett & Duvall, 2004, p. 7), and sought to move beyond singular state-centrism to better understand the power relationships and conflicts that emergent twentieth century (often corporate, private, or non-state) governance structures could create or enforce. As Stoker (1998, p. 17) put it, governance entails ‘creating the conditions for ordered rule and collective action.’ Therefore, it is more than just a capacity, but a specific and complex network of interactions spanning different actors and behaviours.

Some academics studying online life have implicitly adopted similar understandings of governance. As Grimmelmann (2015, p. 47) has suggested, robust systems of community moderation and management are effectively ‘governance mechanisms’ designed to ‘facilitate cooperation and prevent abuse.’ Digital media scholars argue that content policies, terms of service, algorithms, interfaces, and other socio-technical regimes form the
governance mechanisms of today’s online infrastructures (Plantin, Lagoze, Edwards, & Sandvig, 2018). Platform studies scholars show that platform services can significantly affect and mediate individual behaviour (Bucher & Helmond, 2018); therefore, platforms engage in governance at the individual, user level (Gillespie, 2015). But these governance mechanisms are themselves shaped by the policy and regulatory constraints on the corporate entities which deploy these platforms. For a company like Facebook, this includes a host of American regulatory frameworks, international regulatory frameworks for overseas operations, voluntary compliance mechanisms like the Global Network Initiative (GNI) principles, industry-wide voluntary partnerships for terrorist content (initiated at by European Commission), and countless others. Platform governance is an approach necessitating an understanding of technical systems (platforms) and an appreciation for the inherently global arena within which these platform companies’ function. It acknowledges that, as digital media scholars have noted (Gillespie, 2018a), platforms are fundamentally political actors that make important political decisions while engineering what has become the global infrastructure of free expression; but it also acknowledges the other half of the equation: that these private ‘governors’ (Klonick, 2017) are themselves subject to governance on all fronts, and that their conduct of governance is directly informed by local, national, and supranational mechanisms of governance.

The key actors in platform governance therefore include not only users and platform companies, what Poell, Van Dijck, and Nieborg (2018) have called ‘complementors’ (the host of data-brokers, advertisers, developers, and other parties that participate in a platform’s ecosystem), but also, crucially, political actors including various branches of government, as well as other stakeholders and advocacy groups (non-governmental privacy and digital rights groups, academics and researchers, and investigative journalists, who all play a growing accountability function by scrutinizing the practices of platform companies). This is not to say that these political forces (such as state preferences) reign supreme; rather, I suggest that, following the insights of global governance scholarship, ‘a wide variety of forms of governance exist next to each other and that a hierarchy among these various mechanisms is hard, if not impossible, to discern’ (Dingwerth & Pattberg, 2006, p. 192).

Acknowledging media and communications scholars that have done such vital work in advancing the sociological, anthropological, and political economic dimensions of platforms, I posit that to truly understand the power relationships and governance structures underpinning contemporary forms of ‘platform capitalism’ (Šrnicek, 2016), one must also engage with the host of political forces and political (f)actors affecting the platform ecosystem in a variety of ways. As platform politics are becoming increasingly difficult to separate from global politics, an exploration of this kind of platform governance must build on valuable scholarship conducted in various areas, including research from digital media and internet studies, platform studies, political communication, technology policy and law, as well as political science and international relations. Here, I survey insights from this scholarship, grouping them under three arguments: platforms govern, platforms are governed, and platform companies are companies.

**Platforms govern**
Scholars of the internet and digital media have been at the forefront of research into online life, from the initial emergence of ‘social networking sites’ (boyd & Ellison, 2007) to their
eventual transition into ‘platforms’ (Helmond, 2015). The platform studies and game studies traditions within digital media research (Bogost & Montfort, 2007, 2009; Jones & Thiruvathukal, 2012), which primarily view platforms from an architectural perspective (Weltevrede & Borra, 2016), have been bolstered by scholars interested in the power relationships manifest in online cultural practices. Scholars interested in ‘platform politics’ were perhaps the first to focus on governance and power relationships as mediated through platforms. By examining the way in which platform companies navigated the complex, and oftentimes contrasting interests of various stakeholders (including users, advertisers, and regulators), and showing how the term ‘platform’ had itself become a discursive, political imaginary, Gillespie set the stage for work investigating how platforms ‘intervene’ in everyday life, shaping the online experience and algorithmically determining what information to make (in)visible (Bucher, 2018; Gillespie, 2010, 2015). Similarly, Van Dijck (2013, p. 104) specifically discussed how different governance mechanisms affected ‘online sociality’ as experienced by users of Twitter, Flickr, and other services. As researchers explored how best to theorize platforms and situate them critically within past scholarship on social networks, forms of cultural and epistemic power became the focus of researchers such as Langlois (2013, p. 93), who explores platforms as ‘participatory media assemblages, whereby Facebook and Google become conduits of governance.’

But even before the rise of platforms, work on ‘search politics’ and ‘epistemic power’ grappled with what are now the foundational questions of platform governance: how search engines (and the companies that deploy them) may shape knowledge and meaning, and thereby politics, society, and culture (Hargittai, 2007; Introna & Nissenbaum, 2000). A group of scholars that one could call an ‘Amsterdam School’ of critical platform studies provides key insights into platformization, defined as the ‘penetration of economic, governmental, and infrastructural extensions of digital platforms’ into cultural practices (Nieborg & Poell, 2018, p. 2). This approach focuses on ecosystems, and the potentially powerful gatekeeping roles that platforms play by mediating relationships between various parties in the ‘platform society’ (Van Dijck, Poell, & de Waal, 2018, p. 5). Related work by these scholars has mapped the assortment of communicative affordances displayed by different platforms, and how they enable and constrain forms of user behaviour (Bucher & Helmond, 2018; Weltevrede & Borra, 2016). As Lessig (2006, p. 1) famously observed, ‘code is law,’ and the design decisions made by the creator of an online service effectively amount to a form of regulation.

The political implications of the algorithmic systems deployed by platform companies are also of interest to digital media researchers, who, through ‘critical algorithm studies’ scholarship (Gillespie & Seaver, 2015), have examined the increasing role that automated decision-making plays in contemporary life (Beer, 2017; Burrell, 2016). This literature intersects significantly with the current public discourses around platform companies: as Caplan and boyd (2018, p. 2) note, ‘conversations around algorithmic accountability often center on Facebook and Google.’ Although this initially may seem surprising (given the multiple other areas of public life now assailed by artificial intelligence or ‘AI’), it must be recognized that platform companies deploy what are likely the largest, most global, and most widely used algorithmic systems in existence. These systems, due to their scale and (generally) public-facing nature, provide some outcomes that may be relatively visible: while one may be uncertain about when one has been discriminated against by an automated hiring classifier used by a potential employer, problematic
YouTube autocomplete results or racist image tagging systems can provide public examples of bias and discrimination. Therefore, work striving to understand how to ‘govern algorithms’ (Barocas, Hood, & Ziewitz, 2013; Ziewitz, 2016) fundamentally implicates platforms. After all: how can we strive for accountable algorithms, if the corporate entities that build and deploy them are not fair, accountable, transparent, or ethical, and if they seem to be entrenching, rather than combatting, existing social prejudices?

While this scholarship has illustrated how platforms govern users through their design, architectures, assemblages of algorithms, and other technical structures, the mounting explorations of how platforms interact with social structures more broadly is also vital. Intersectional arguments that online services can encode gender dynamics, class structures, and racism (Noble, 2018) provide an important depiction of how platform companies can engage in governance at a broader level. How do the decisions made largely by a homogenous group of white elites in Silicon Valley affect different users around the world? Nakamura, Bivens, and others have demonstrated how the choice of architectures and other design decisions (such as sign-in pages that present new users with a binary gender option) can entrench normative judgements around gender, race, class, and sexuality (Bivens, 2017; Nakamura, 2013). By adopting a sociotechnical perspective, this scholarship highlights the importance of internal dynamics within platform companies, and of employees and organizational structures within platform governance (e.g., the lack of diversity within engineering teams has a major impact on the design decisions those teams make).

Work that scrutinizes the specific practices of platform companies provides another important contribution. Political communication researchers have assessed the growing role of technology firms as ‘active agents in political processes’ through their direct collaboration with campaigns (Kreiss & Mcgregor, 2018, p. 155), and their troubled interactions with publishers and news organizations (Bell, Owen, Brown, Hauka, & Rashidian, 2017; Nielsen & Ganter, 2017). A growing body of work seeks to explore the practices and implications of commercial content moderation and content policy (Roberts, 2018) as the processes through which platform companies set and enforce the rules governing user speech (Gillespie, 2018a; Suzor, 2019). Researchers have thoughtfully documented both how this culminates in governance at the user level and how users themselves perceive and interact with these structures (Duguay, Burgess, & Suzor, 2018; Myers West, 2018). This work builds on research in human–computer interaction and community management, where ‘platform governance’ refers to the systems of rules, norms, and civic labour governing an online community (Matias & Mou, 2018). Digital media, internet studies, and communication scholarship forms the base of a platform governance approach: it provides an appreciation for the functions, affordances, and politics of contemporary platforms, and illustrates the contours of how platform companies currently govern user behaviour. However, it could be better contextualized within the broader context of platform companies as corporate actors, and the contested global governance arena within which they operate.

**Platform companies are companies**

Media scholars have called for more involvement from political scientists in contemporary debates around platform companies. As Helberger et al. (2018, p. 4) note, ‘the governance of platforms and how they are involved in the organization of societal relationships is ultimately a political question,’ but political scientists have yet to contribute substantially to
current debates around democracy, politics, and platform companies. One of the simplest, yet potentially most impactful insights that can be provided to platform studies research from political science is that platform companies may span industries and resist meaningful classification, but they are still companies. Platform governance work should thus draw on the substantial scholarship which has examined the role of influential global corporations in public life and public affairs around the world.

As platform companies become transnational corporations perhaps unmatched in their global reach and wealth (Apple recently became the first company in history to obtain a market capitalization of more than a trillion dollars), a clear point of reference is historical research about the political influence of other global corporations. As Moore and Tambini (2018) argue, there are some parallels that can be drawn between the current ‘digital giants’ and monopolistic firms like Standard Oil, which played a significant economic and political role the nineteenth century United States. But there are further parallels to the post-WW2 period, where the emergence of the international order and the slow advent of globalization allowed a group of firms to grow into multinational corporations of unprecedented wealth and size (Vernon, 1977). As political economists such as Strange (1991, 1996) observed, the corporations that drove economic growth and boosted national gross domestic products also created many governance challenges, testing jurisdictions and traditional forms of regulation. Keeping corporations accountable became a pressing global governance problem as mounting evidence of their repeated evasion of labour and environmental standards by large corporations – especially extractive natural-resource based multinationals – emerged in the 1970s and 80s (Keck & Sikkink, 2014; Ruggie, 2008).

The scholarship that arose to catalogue the relationships between global corporations and other actors in domestic and international politics (Fuchs, 2007; Mikler, 2018) is another important area which could therefore contribute to future platform governance research. Despite their constant invocation of the rhetoric of disruption and innovation, platform companies function in many ways as traditional corporate actors. They are tremendous lobbyists: in 2017, Google spent more on lobbying in Washington than any other company (Shaban, 2018), and Facebook has now on multiple occasions hired lobbying firms to help discredit their competitors (Nicas & Rosenberg, 2018). They minimize their tax burden with classic profit shifting techniques, and they deploy contractors to keep their workforces small and relatively inexpensive (Srnicek, 2016). As corporations, platforms can be therefore governed along the lines of traditional multinational/global enterprises.

Decades ago, global advocacy groups organized against firms like Nestlé, ‘naming and shaming’ them into more socially responsible business practices (Sikkink, 1986). International organizations built soft forms of governance through codes of conduct and Corporate Social Responsibility networks (Ruggie, 2013). Activist shareholders sought to reform companies from within, and employees within the firms made their voices heard to push for change. There are clear parallels to today, where international digital rights organizations (e.g., Privacy International), investigative journalists (e.g., ProPublica), and academics have become key actors creating public pressure for responsible platform governance. Furthermore, non-governmental ‘social responsibility’ mechanisms, first created for freedom of expression issues (such as the GNI) are now potentially being revitalized for the broader host of concerns that platform companies need to deal
with today (Kaye, 2018). These processes will be contested, and provide no panacea, as keeping global corporations accountable is no easy task (Ruggie, 2008). However, future platform governance scholarship can learn from the past as it seeks to devise new forms of corporate accountability suitable for the business models of today’s data-driven platform corporations, and as it evaluates the emerging international structures, processes, or organizations that can help fill loopholes in how platform companies are governed by national legislation.

**Platforms are governed**

Governance by platforms does not happen in a vacuum. Although it has become commonplace to argue that platform companies are profoundly under-regulated, and that platforms wield almost unfettered, sovereign power at global scale (Vaidhyanathan, 2018), the current regulatory landscape across the Global North (laissez-faire as it generally is) can be interpreted as the product of American foreign policy interests (Goldsmith, 2018). It may be true that efforts to reform platform companies in areas such as privacy have yielded mixed results, especially in the United States; however, the efficacy of a specific regulatory intervention does not change the reality that platforms can be and are governed by governments. The practices of platform companies are subject to local laws (e.g., platform companies heed governmental notice-and-takedown requests on content), and can be scrutinized by regulators, privacy authorities, competition authorities, and other institutional actors at the domestic or international level.

Therefore, legal scholarship is instrumental in helping to formulate an understanding of how the behaviour of platforms, as online intermediaries enabling activity by third-parties, is governed externally by telecommunications and internet law (Goldsmith & Wu, 2006). In the United States, this includes the study of intermediary liability provisions, such as Section 230 of the Communications Decency Act, and the implications of its reform (Citron & Wittes, 2017), as well as comparative, international explorations of how differing liability frameworks create different incentives for platform behaviour (Gasser & Schulz, 2015; Omer, 2014). Media law and information policy provides another perspective into how platforms may fall under existing regulatory structures as publishers and information conduits (Brown & Marsden, 2015; Lunt & Livingstone, 2011).

Alongside the ‘soft’ international governance and accountability structures discussed in the previous section, platform companies are also potentially governed internally. While technology corporations often have shareholder structures that consolidating power within their executive (preventing them from being fully accountable to their investors), it is important to remember that they are complex collections of individuals and interests as opposed to unitary actors. Employees (either full-time employees, or the ‘gig’ workers that many platforms rely on) can unionize, organize for responsible product design, stage protests and walkouts, and try and steer a company’s actions from the inside, either collectively or individually (Wood, Lehdonvirta, & Graham, 2018).

**Emerging governance modes**

In the decade since the Arab Spring and the various ‘Twitter Revolutions,’ the political role of platform companies has been increasingly problematized. Our understanding of how platforms govern (via content policy decisions, algorithms, and forms of user
‘dependency’) (Nieborg & Poell, 2018) has deepened considerably. But the question of how platforms should be governed going forward – currently being discussed in many policy and public circles, though less so in scholarly ones – remains open, even as regulatory options are being exercised by decision-makers around the world. These conversations are taking on an air of inevitability: as Pasquale (2018, p. 14) puts it, ‘even Mark Zuckerberg recently conceded that the question is not whether to regulate Facebook, but how.’ It is the latter question that will prove essential moving forward. If users of the services provided by platform companies seek ‘good governance’ – in other words, seek to recapture what was once heralded by the open Web – than how should they get it? What are the policy steps that should be taken? And what should be the first principles underpinning those steps? Answering these questions will involve finding the appropriate balance between the ‘self-governance,’ ‘external governance,’ and ‘co-governance’ models that are currently developing.

**Self-Governance**

The current dominant governance mode is often referred to as ‘self-governance’ or ‘self-regulation.’ This approach, enshrined through legislation like the US Communications Decency Act and the EU E-Commerce Directive, limits platform liability and results in a relatively laissez-faire relationship between governing institutions and platform companies. Today, companies own and operate what is often highly visible, highly trafficked ‘public’ space, and respond to third-party complaints about content (for reasons ranging from intellectual property to national security). The companies are generally not liable for what users do on a platform, as long as they take adequate steps to redress third party ‘notice.’ In this governing mode, transparency is generally voluntary, and most platform decisions are made with minimal external oversight (Suzor, 2019).

Since 2016, platform companies have implemented multiple changes in response to public concern. These initiatives, which range from new advertising tools to changes as to how they interact with political campaigns, seem designed to head off possible avenues of regulation while also effectively maintaining the highly-profitable status quo. In the past two years, these self-regulatory improvements have prominently consisted of technical changes or tools, transparency efforts, or some combination of the two (Garton Ash, Gorwa, & Metaxa, 2019).

Platform companies insist that they can be accountable to their users by slowly increasing transparency in numerous areas, such as content policy and advertising. After a group of US Senators proposed the Honest Ads Act, Facebook took many of the provisions of the act and pre-emptively implemented them (Timmons & Kozlowska, 2018). Facebook, Google, and Twitter now require political advertisers to register or provide identification in certain jurisdictions, and both Facebook and Google have built public-facing tools where researchers or interested members of the public can see ads that are being deployed, along with some information about who is paying for them, and how much (Gorwa & Garton Ash, Forthcoming). Other major transparency initiatives have been launched as part of this broader effort to regain public trust. In April 2018, Facebook made the important step of releasing public-facing internal guidelines for their ‘Community Standards,’ the rules that govern what the more than 2.2 billion monthly active users of Facebook can post on the site (Bikert, 2018). They also launched an ongoing project partnering
with academics which hopes to create a reputable mechanism for third party data access and independent research (King & Persily, 2018).

As these changes illustrate, the self-governance model has numerous advantages. Through investigatory journalism, academic engagement, and public advocacy, companies can be nudged in the right direction without complex regulatory interventions (an especially challenging prospect when, due to corporate secrecy and the inherently ‘black box’ nature of contemporary platform companies, the true scope of many of the problems in today’s platform ecosystem are not fully known). Platform companies can quickly make specific interventions (such as requiring advertisers to register their identities or combating hate speech through specific tweaks of their content policies) far before legislation goes into effect. Additionally, by keeping key decisions about free expression largely in the hands of online intermediaries, important concerns about government censorship and suppression (including in less democratic countries that may be keen to exert control over online environments) are assuaged (Kaye, 2018; Keller, 2018). But there are also important limitations: voluntary arrangements rely on goodwill and provide limited recourse in the case of non-compliance. Many recent transparency initiatives are predominantly public-facing, providing useful tools for journalists and interested members of the public, but arguably much less useful information for regulators and investigators. Furthermore, these minor changes do little to provoke systemic change or modify platform business models, which may be fundamentally problematic and based on extractive surveillance and data collection (Fuchs, 2012; Zuboff, 2015).

**External governance**

Following countless public relations scandals, large-scale privacy breaches, and growing concerns about polarization and misinformation, many are advocating for more ‘external governance.’ In practice, government intervention is catalyzing around three policy levers: the implementation of comprehensive privacy and data protection regulation, the repudiation of intermediary liability protections, and the use of competition and monopoly law. Softer versions consist of specifically crafted legislation around existing policy levers (e.g., amending existing electoral laws with online advertising provisions). New forms of legislation targeted towards platforms demonstrate the most discrete effort to affect the business models and practices of certain companies and industries. The German Network Enforcement (NetzDG) law, born out of apparent frustration with the self-governance mode (with German officials arguing that American platform companies like Twitter and Facebook were responding too slowly or incompletely to their requests), removes liability protections for content violating German law, mandating that (sufficiently large, aka non-German) platform companies remove ‘evidently unlawful’ material in less than 24 h following a complaint (Schulz, 2018, p. 6), or face significant fines. NetzDG also seems to be the first legislation to move beyond current voluntary transparency measures by legally mandating comprehensive transparency reporting on operations in Germany (Keller, Forthcoming).

Privacy legislation is another lever: under the European Union’s 2016 General Data Protection Regulation (GDPR; in effect as of May 2018), platforms are given clear requirements as to how they process personal data, with the threat of enormous penalties for non-compliance (Golla, 2017). GDPR has other major stipulations about data portability, data
protection by design, and informed consent that, if enforced, could have a resounding impact on the platform ecosystem (Hildebrandt, 2018). Although it remains unlikely in the short-term, it is increasingly being argued that the United States should adopt similarly comprehensive legislation for privacy and data rights (Tisne, 2018). Perhaps the most drastic measure, however, would involve the use of anti-trust law in the United States. Although platform companies have already faced fines from European competition authorities, some legal scholars have argued that these companies should be broken up or prevented from making future acquisitions (Pasquale, 2018; Wu, 2018). From this perspective, the argument is that, despite their provision of ostensibly ‘free’ services (and therefore, the lack of clear price discrimination), users are harmed by multiple forms of anti-competitive behaviour exhibited by platform companies (Khan, 2017).

**Co-governance**

Steps towards ‘co-governance’ seek a third way between the two previous approaches. In the short term, such models seek to provide some values of democratic accountability without making extreme changes to the status quo. Civil society organizations, for example, have advocated for some kind of organization that could perform multiple functions ranging from investigating user complaints to creating ethical frameworks for platform companies, perhaps modelled after international press councils which set codes of conduct and standards for news organizations (ARTICLE 19, 2018). Kaye (2018, p. 18), describes the possibility of various ‘ombudsman programmes or third-party adjudication’ systems to which users could address complaints and seek redress, and outlines several historical organizations which could provide precedent. Much as the GNI brought an international group of civil society organizations, academics, and other stakeholders together with platform companies to establish best practices for the promotion of free expression (and a system of transparency reporting, third party audits, and other mechanisms to help oversee those practices), a similar organization could be formed to tackle recent concerns around disinformation, hate speech, privacy, and more. One of the major developments since the GNI is not just that the platform’s external conduct vis-à-vis governments must be scrutinized via audits and other accountability mechanisms. Today, the internal behaviour of the platforms themselves is of key public interest and demands increased oversight and stakeholder engagement.

Other possible visions of co-governance involve more granular forms of user participation in policy decisions: for instance, Gillespie (2018a) describes Facebook’s brief experiment with user voting on policy changes as a possible model which need not have been cast aside so quickly. Will platforms seek to bring in external community support into their governance practices in order to gain legitimacy and trust? In November 2018, Zuckerberg announced that the French government would be permitted to embed regulators into content policy processes in the country, and that Facebook would create a ‘supreme court’ that would allow external appeal for content policy decisions (Zuckerberg, 2018). While the outcome of these changes is yet to be seen, they signal an understanding on Zuckerberg’s behalf that self-governance is no longer seen as a satisfactory long-term solution by various governance stakeholders. Co-governance could also provide even more radical options in the long term, as it lends itself to a philosophy that leads away from major, corporatized platforms and towards various platform cooperatives, decentralized
systems, and other forms of community self-management. While these remain admittedly unlikely in the near-term (due to scale issues, network effects, and other challenges), they could provide a more equitable and just digital economy for users in the long term (Scholz, 2016; Scholz & Schneider, 2017).

Conclusion: normative principles for good platform governance

Given the rapid pace and development of the platform ecosystem, as well as the dynamic nature of the platform companies in question, new models for digital governance will likely need to be developed. As Nooren, van Gorp, van Eijk, and Fathaigh (2018, p. 282) suggest, when it comes to platforms, ‘a more normative/functional approach is required instead of the overly detailed regulation common to static markets … it may be necessary to move toward “principles-based regulation” as opposed to “rules-based regulation.”’ A platform governance approach should therefore seek not only to understand the complex governance relationships and public policy challenges in today’s platform society (Nash, Bright, Margetts, & Lehdonvirta, 2017), but also ask how these relationships can become more beneficial for the multitude rather than the few.

Recent research has begun extending some possible guiding principles, each with their relative merits and challenges. These include rights-based legal approaches, such as international human rights law (Kaye, 2018; Suzor, Van Geelen, & Myers West, 2018) or American civil rights law (Citron, 2009), which could possibly provide an avenue for platforms to better ground policy decisions and build legitimacy. Other possible frameworks include aspirational computer science principles for algorithmic systems, including fairness, accountability, and transparency, ethics, and responsibility (Diakopoulos et al., 2016); the developing area of ‘data justice,’ which seeks to transcend the focus on individual harms wrought by various data-driven systems and instead apply holistic principles of social justice at a broader level (Dencik, Hintz, & Cable, 2016; Taylor, 2017). Others could include political science and governance studies principles of meaningful democratic accountability and transparency (Fung, 2013; Hood & Heald, 2006), or Corporate Social Responsibility and other past efforts to steer business in the direction of human rights or other normative aims (Urban, 2014). How can these various values be enshrined through sensible legislation and accountability mechanisms? Principles, values, and imaginaries will be crucial for grounding the future platform governance research agenda, and scholars will need to be assertive with their ideas and their output. Change will not come easily: platform companies have become influential political actors with an obvious interest in preserving their dominant market positions. Creative ideas will be needed to help ‘disrupt’ the ‘disruptors’ and introduce fairer, more accountable, and more just forms of platform governance.

Notes

1. For instance, past efforts to foist meaningful oversight mechanisms onto platform companies, such as the Federal Trade Commission’s 2011 and 2012 consent decrees regarding Google and Facebook’s deceptive privacy practices are seen to have largely failed (Gray, 2018).
2. An excellent example of this shift is seen in Emmanuel Macron’s speech at the 2018 Internet Governance Forum in Paris.
Acknowledgements

Many thanks to the participants in the #AoIR2018 session on ‘Platform Governance and Moderation’ for the feedback and critique that helped shape this article from the onset, and to the special issue editors, Alison Harvey and Mary Elizabeth Luka, for engaging so thoughtfully and carefully with not only this article’s arguments but also its structure and style. I am further indebted to Nicolas Suzor, Corinne Cath, Thomas Poell, Ralph Schroeder, and the anonymous reviewers of ICS for taking the time to read and provide excellent comments.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

The author would like to thank the Social Science and Humanities Research Council of Canada (SSHRC) and the Canadian Centennial Scholarship Fund for supporting his studies. As well, the author gratefully acknowledges travel support from the European Research Council grant, “Computational Propaganda: Investigating the Impact of Algorithms and Bots on Political Discourse in Europe,” Proposal 648311, 2015-2020, Philip N. Howard, Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the European Research Council.

Notes on contributor

Robert Gorwa is a doctoral candidate in the Department of Politics and International Relations, University of Oxford, where his dissertation examines the political role of large technology ‘platforms,’ with a focus on evolving notions of corporate power and private governance in contemporary democracies. A graduate of the Oxford Internet Institute (MSc, 2017), and a researcher affiliated with Oxford’s Centre for Technology and Global Affairs and Stanford’s Project on Democracy and the Internet, his writing on technology and politics has been published in Foreign Affairs, Wired Magazine (UK), the Los Angeles Review of Books, the Washington Post, and other popular outlets.

robert.gorwa@politics.ox.ac.uk

ORCID

Robert Gorwa http://orcid.org/0000-0002-4891-5053

References


By Evelyn Douek  Thursday, April 26, 2018, 7:00 AM

DayZero: Cybersecurity Law and Policy

During Mark Zuckerberg’s 10-plus hours of testimony before Congress on April 10 and 11, the Facebook CEO was asked at least six times about his company’s censorship of conservative internet personalities Diamond and Silk—to the point where New York Times technology reporter Kevin Roose commented, “I think if you were an alien who dropped down to earth to observe the testimony ... you would think there was no one on earth who was more important than Diamond and Silk.” In contrast, Zuckerberg fielded exactly three questions about his company’s role in the ongoing violence in Myanmar—one of the most pressing human rights atrocities of our time.

This was a missed opportunity. At this point, that Facebook has played a role in spreading the violence in Myanmar is indisputable. The ongoing crisis raises many legal questions and exposes a lacuna in the law that political and public pressure is currently trying to fill. But one thing is clear: This crisis should be an important part of the conversation about the role of the big technology companies in societies around the world. If developed countries are concerned with the impacts of social media on civil discourse, they should not forget that the same impacts can have more drastic effects in developing countries—and will continue to do so as millions more people come online.

Ethnic Cleansing in Myanmar

The U.N. High Commissioner for Human Rights, Zeid Ra’ad Al Hussein, has called the latest wave of violence against Rohingya Muslims in Myanmar has a “textbook example of ethnic cleansing.” The crackdown by the country’s security forces, which began in August 2017, has precipitated the fastest-growing refugee crisis in the world: Over 80,000 Rohingya refugees in total have now fled to Bangladesh. Myanmar has blocked most international access to Rakhine state, where the violence is occurring, making reliable estimates of death tolls near impossible. Doctors Without Borders said late last year that its conservative estimates of casualties were at least 900, including at least 30 children under the age of five.

This is just the latest flare-up of long-simmering ethnic tensions between the country’s Rohingya Muslim minority and its Buddhist majority, who view the Rohingya as illegal immigrants from Bangladesh. Zeid has drawn attention to the progressive stripping of the political and civil rights of the Rohingya population by “successive Myanmar governments … since 196.” But the tension has come to the fore again since the country started its transition to democracy in 2010, as deep ethnic tensions have erupted into periodic waves of violence and fractured a society no longer suppressed by military rule.

Facebook in Myanmar

Facebook founded the controversial “Free Basics” initiative in 2013 under the name Internet.org, with the goal of bringing greater internet connectivity to rural and low-income populations around the world. The Free Basics platform allows users to access supported services for free by giving them a “zero-rating,” meaning that those services did not count towards the user’s data cap. Free Basics launched in Myanmar in 2016, and its influence has been profound. Free Basics’ arrival coincided with the arrival of the internet more generally: Internet penetration in Myanmar has gone from only 2 percent in 2013 to 25.1 percent in 2017 and over 90 percent of the country’s population now have access to a phone with internet service.

This means that to many in Myanmar, “Facebook is the internet.” Facebook entered into a society that had low internet literacy to begin with, carrying the characteristics of its platform that have also caused controversy elsewhere: promotion of echo chambers that cause fragmentation and polarization of the public sphere, algorithms that optimize for engagement and prioritize extremist content, and the capacity for information to go “viral” and reach audiences at unprecedented speed and scale.

As a result, Facebook has become a primary conduit for the spread of anti-Rohingya propaganda and hate speech. U.N. investigators have directly pointed the finger at Facebook as playing a “determining role” in the violence, saying that “Facebook has now turned into a beast, and not what it originally intended.” Reporting has shown that hate speech spreads through “affinity groups” in Facebook or WhatsApp group messages that include friends or relatives who are “trusted sources.” As people are often members of many such groups, hate speech and misinformation jumps from group to group, each time appearing to come from a trusted source.
Facebook has also given key figures a megaphone. The leader of a prominent group of radical Buddhist monks, Ashin Wirathu, built up a large following on the platform before Facebook permanently disabled his account in January this year. A New York Times reporter in Myanmar, Paul Mozur, said Wirathu used to print paper flyers to spread his messages, but Facebook allowed him to get 100 times the reach. And it is the singular feature of the internet that Wirathu could also obtain that reach more quickly and cheaply online.

_latfooted Facebook_

In 2013, almost five years ago, Zeynep Tufekci, a leading researcher on social implications of digital connectivity, tweeted:

```
zeynep tufekci

last year ,lpndered the fist socialedia f ueled ethnic cleansing,fM ethnic tensions expanding online hate spech,is it Burra?

2 5/1 2
```

That is to say: The recent events were not only foreseeable, but they were actually foreseen. Reporting at the time documented the proliferation of hate speech on the platform. A 2014 New York Times article on violent riots that left two dead described how the riots were “set off” by unconfirmed rumours of a rape on Facebook and noted the influence of Wirathu’s radical teachings.

And yet Facebook has largely been an “absentee landlord” in Myanmar, as Phil Robertson, deputy director of Human Rights Watch in Asia, has put it. Facebook has no office in Myanmar, which makes it much more difficult for the company to evaluate the complicated factors inherent in content moderation decisions that require deep understanding of context. Facebook does not disclose how many content moderators it has with local language expertise—those it does have appear to work in Dublin, where it apparently has some difficulty finding qualified candidates.

During the congressional hearings earlier this month, Zuckerberg was prepared to answer questions about his company’s role in Myanmar with a three-step plan: hiring dozens more Burmese-language content reviewers, working with civil society in Myanmar to take down accounts of “specific hate figures,” and creating a product team to implement (unspecified) product changes in Myanmar and other countries that may have similar issues in the future.

But Zuckerberg announced these steps only after a large amount of recent press coverage of the situation. Even as recently as March 15 this year, Facebook didn’t seem to have a plan. In an interview with Slate, Facebook’s head of newsfeed, Adam Mosseri, said that while people at the company “lose some sleep over this,” the situation was particularly challenging for a number of reasons—including that there were no local third-party fact-checkers for Facebook to partner with. Civil society organizations in Myanmar have capitalized on public attention, publishing an open letter specifying the problems of Facebook’s content moderation in the country, in particular:

- an over-reliance on third parties (to flag content, if they came across it in time, rather than monitoring for such content itself);
- a lack of proper emergency escalation mechanism (it had taken days for Facebook to step in after the organizations had tried to raise concerns about the messages, and they went viral in the meantime);
- lack of engagement with local stakeholders (requests to talk to Facebook’s engineering and data teams about systemic solutions had gone unanswered); and
- a lack of transparency (seven months after an incident Zuckerberg cited as a success story because Facebook had blocked a series of messages inciting specific violence, the organizations still did not know the details of what had happened).

The overall picture that these facts paint is that Facebook’s role as an instrument in ethnic violence in Myanmar is a recurrent and systemic problem, and one which the company has been (or should have been) aware of for years.
Despite this background, Facebook has few binding legal obligations and the relevant legal frameworks are weak.

First, and most obviously, Facebook must comply with the local laws of jurisdictions in which it operates. However, the government of Myanmar has shown little appetite for placing political or legal pressure on Facebook to improve its operations, instead downplaying the level of violence altogether. The government has consistently denied claims that genocide or ethnic cleansing are occurring, claiming that the Rohingya are burning their own villages. Aung San Suu Kyi, the country’s de facto leader—and the recipient of a Nobel Peace Prize of which there have been calls for her to be stripped given her inadequate response to the crisis—has even called reports of the crimes “fake news.” Her office has itself used Facebook to disseminate inflammatory content contradicted by reports from local journalists.

International law in this context also lacks teeth. Myanmar is one of the 26 states that has not ratified the International Covenant on Civil and Political Rights, and therefore has not committed itself to the obligation to ensure that corporations respect the human rights of people within its territory. It is also not a state party to the Rome Statute, which makes it hard (absent a U.N. Security Council referral) for International Criminal Court prosecutors to bring charges for crimes against humanity. (However, a filing published earlier this month argues that the ICC has jurisdiction because the crime of deportation is occurring on the territory of Bangladesh, a state party.)

Of course, Facebook is also not a party to these conventions: It is not a state, despite growing comparisons to that effect. As such, these international human rights instruments do not bind it directly. In some ways, then, the Myanmar crisis is just another incarnation of the difficulty international law has had in holding multinational corporations accountable for the impacts of their operations on rights. The U.N. Guiding Principles on Business and Human Rights were developed in order to help bridge this gap. But the principles are non-binding, and Facebook is not legally liable if it fails to adopt them.

The Guiding Principles do, at least, create a framework for thinking about Facebook’s responsibilities—including its duty to not contribute to human rights abuses, conduct due diligence on its human rights obligations and provide transparency into its impacts. These obligations are reflected in the complaints of the Burmese civil society organizations in their open letter, and Zuckerberg’s commitments to Congress would go some way towards meeting them. “Dozens” more Burmese language reviewers will help, but in comparison, Facebook said it would hire an additional 3,000 moderators by the end of 2018 and open a second content moderation center in Germany when the country passed laws threatening large fines if hate speech wasn’t removed within 24 hours. This is a stark differential, and one that seems unlikely to be wholly justified by reference to the estimated number of users in each country alone (14 million in Myanmar, compared with 39 million in Germany in 2017). Furthermore, these newfound commitments suggest systems have been inadequate for the intervening years, and underline the necessity of monitoring their effective implementation.

Moreover, there are some features of Facebook’s business that make it a more difficult case than the typical multinational corporation. It operates in many countries without any physical presence at all—meaning that in states in which the government does want to regulate Facebook, its options and leverage are much more limited. In Sri Lanka, for example, officials found Facebook unresponsive to official concerns about hate speech and misinformation on its platform until the government blocked access entirely. The head of public information told the New York Times that big companies like Facebook “look at us only as markets ... We’re a society, we’re not just a market.” This is the second way Facebook is distinguishable from other multinational corporations—its dominance in these markets means it becomes the de facto public sphere. It is not just like any other commercial product: It is integral to the flow of information.

Finally, Facebook’s global reach means that these individual markets are insignificant to the company as a whole. Facebook has become integral to Myanmar, but Myanmar matters very little to Facebook.

Abader Conversation

It’s important to avoid the urge to oversimplify this issue. Myanmar’s example raises difficult questions, factually, legally, and philosophically.

The exact nature and extent of Facebook’s contribution to the spread of violence in Myanmar is unknown and perhaps unknowable. The connection between hate speech and mass atrocities does have a long and bloody history in the 20th century alone. Nazi Julius Streicher’s anti-Semitic articles in Der Führer were found by the Nuremberg Tribunal to have “infected the
German mind with the virus of anti-Semitism and incited the German people to active persecution.” “Hate radio” played a key role in the Rwandan genocide, and villages with access to radio broadcasts had increased participation in the killings. As Samantha Power has written, “[k]illers often carried a machete in one hand and a transistor radio in the other.”

So U.N. investigators have precedent of declaring that the spread of hate speech through Facebook has had a “determining role” in the violence in Myanmar. Academic studies are shedding light on the correlation, as is journalism. In a recent article about recent outbreaks of violence in Sri Lanka, Amanda Taub and Max Fisher of the New York Times “found that Facebook’s newsfeed played a central role in nearly every step from rumor to killing. Facebook officials, [interviewees said], ignored repeated warnings of the potential for violence, resisting pressure to hire moderators or establish emergency points of contact.” But correlation is not causation, factually or legally. As Susan Benesch has written, “[t]he idea that inflammatory speech is a catalyst for genocide is widely believed, likely correct and of no small importance ... But the impact of speech on groups is complex, and difficult to measure or prove.”

Ethnic violence existed in Myanmar long before Facebook arrived, and Facebook is hardly the most culpable actor in the spread of misinformation and hate speech. Facebook’s mission to “bring more people online and help improve their lives” is not a fig leaf but a noble goal, in many cases realized. People in Myanmar rely on Facebook for much of their communication, as do news outlets and journalists for getting important information out. The platform is becoming increasingly important for local businesses. Even if it were possible, it is by no means clear that Myanmar would be better off if Facebook exited the country tomorrow.

These complexities should be borne in mind when thinking about how a company like Facebook could be more regulated internationally. As Tufekci sums it up,

---

There are questions to ask about Facebook’s flat-footedness in the face of clear evidence of its role in a mass atrocity. Myanmar is only the most vivid example of dynamics playing out in many developing countries—and as developed countries with political and economic power brainstorm ideas for the regulation of Big Tech, this should be a part of the calculation. Legal frameworks have the power to provide standards against which Facebook’s conduct could be measured, and a common vocabulary with which to have the conversation. They could also provide procedural mechanisms that provide transparency into what resources Facebook is dedicating to minimizing its contribution to harm in these situations and visibility into specific issues such as the steps it is taking to prevent moderators from again removing content that could potentially be important evidence of human rights abuses.
International law will need to account for the role of new communications technologies in crimes against humanity, and the unprecedented power a few corporations have over the public spheres of foreign countries. But this reckoning will take time, and meanwhile, domestic legal systems are already contemplating action. As they do so, lawmakers in countries like the United States should consider what tools are available to shed light not only on the effects of the activities they are thinking of regulating domestically, but also on the power of these platforms in societies with weaker institutions.

To ask Facebook to remove all hate speech or misinformation from its platform is unrealistic and unhelpful. But to ask it to take reasonable steps to create systems to prevent the spread of the majority of the worst content in a timely fashion could save more than a few lives.
YouTube’s Bad Week and the Limitations of Laboratories of Online Governance

By Evelyn Douek     Tuesday, June 11, 2019, 12:05 PM

DayZero: Cybersecurity Law and Policy

The techlash has well and truly arrived on YouTube’s doorstep. On June 3, the New York Times reported on research showing that YouTube’s recommendation algorithm serves up videos of young people to viewers who appear to show sexual interest in children. In any other week this might have been a huge public controversy, but the news was consumed instead by a very different content moderation blow-up. Centering around the meaning of YouTube’s harassment and hate speech policies and whether a right-wing commentator with nearly four million subscribers had violated them, the week-long saga illustrates how different platforms are developing very different approaches to handling high-profile disputes about what they allow on their services.

What Happened?

It started on May 30, when Vox journalist Carlos Maza wrote on Twitter that a right-wing commentator named Steven Crowder had been routinely mocking Maza’s voice and mannerisms and describing him using racist and homophobic language. According to Maza, the harassment had been ongoing for years and included a doxxing where his phone was bombarded with hundreds of texts, all reading “debate steven crowder.” Maza was clear that his complaint was directed not at Crowder but at YouTube, “which claims to support its LGBT creators, and has explicit policies against harassment and bullying …. But YouTube is never going to actually enforce its policies. Because Crowder has 3 million YouTube subscribers, and enforcing their rules would get them accused on anti-conservative bias.”

After Maza’s thread went viral and received coverage from major news outlets, the @TeamYouTube twitter account responded, saying that the company was “looking into it further.” On May 31, Crowder released a video titled “Vox is Trying to Ban This Channel” (although Maza made clear that his tweets were done in his personal capacity and not on behalf of Vox). YouTube’s response finally came on June 4. In a cursory four-part Twitter thread, YouTube wrote, “Our teams spent the last few days conducting an in-depth review of the videos flagged to us, and while we found language that was clearly hurtful, the videos as posted don’t violate our policies.” To “explain” the decision, YouTube said that “[o]pinions can be deeply offensive, but if they don’t violate our policies, they’ll remain on our site.” In an email to a journalist, YouTube clarified that it takes into account whether “criticism is focused primarily on debating the opinions expressed or is solely malicious.” Crowder’s videos, the company stated, did not violate the policy because “the main point of these videos was not to harass or threaten, but rather to respond to the opinion” expressed by Maza.

A day later, amid ongoing backlash, YouTube provided a further single-tweet update, announcing that it was demonetizing Crowder’s channel—that is, removing Crowder’s ability to earn money by running ads—“because a pattern of egregious actions has harmed the broader community.” The company linked to a blog post explaining that in rare cases when a creator does something “particularly blatant … it can cause lasting damage to the community, including viewers, creators and the outside world.” In these circumstances, “we need a broader set of tools at our disposal that can be used more quickly and effectively than the current system of guidelines and strikes.” These tools include demonetization and removal from YouTube’s recommendations.

Yet these actions did little to halt the slow-motion fiasco. Maza pointed out that YouTube demonetization would mean little to Crowder, who also makes money by selling merchandise, including a T-shirt reading “Socialism is for F*gs.” YouTube “clarified” that in order to reinstate monetization, Crowder would need to remove the link to selling those T-shirts. After this position stoked further outrage, YouTube apologized for the “confusion” and retreated to the view that the problem was not the T-shirts per se but, again, the “continued egregious actions that have harmed the broader community.” Meanwhile, as many rallied to support Maza, others took Crowder’s side: Senator Ted Cruz denigrated YouTube as the Star Chamber.

It was not until June 5 that YouTube provided anything like a full explanation for its actions on a forum other than Twitter. In a blog post titled “Taking a harder look at harassment,” the company wrote, “These are important issues and we’d like to provide more details and context than is possible in any one string of tweets.”

YouTube has two key relevant policies: harassment and cyberbullying and hate speech. The policy on harassment reads:
Content or behavior intended to maliciously harass, threaten, or bully others is not allowed on YouTube. …

Don’t post content on YouTube if it fits any of the descriptions noted below. …

- Content that is deliberately posted in order to humiliate someone
- Content that makes hurtful and negative personal comments/videos about another person
- Content that incites others to harass or threaten individuals on or off YouTube

For hate speech, YouTube says it will remove “content promoting ... hatred against individuals” or using stereotypes to promote hatred based on attributes including nationality, race and sexual orientation.

YouTube’s ultimate blog post elaborated on these guidelines. For harassment, the question is whether the purpose of the video, taken as a whole, is to harass or humiliate. For hate speech, the same question: whether the “primary purpose” is to incite hatred. In either case, “using racial, homophobic, or sexist epithets on their own would not necessarily violate either of these policies.” Although it appeared that moments from the supercut Maza originally posted were facially inconsistent with YouTube’s policies, YouTube had decided that, in context and considered as a whole, Crowder’s videos were not harassment or hate speech. But it also promised to take “a hard look at our harassment policies with an aim to update them.”

In an interview on June 10, YouTube CEO Susan Wojcicki explained why the company had made the call to leave the videos up but demonetize them, saying that the company had a “higher standard” for creators who earn money from their videos.

The Meaning of YouTube’s Community Guidelines

The whole controversy centers around the meaning of YouTube’s community guidelines. This is true in two senses. The first is about the operational meaning of YouTube’s policies and whether the Crowder videos violated them. But the second, deeper debate is about what it means for YouTube, a private company, to have these guidelines at all. Whatever your view of the underlying substantive issue of what YouTube should or should not support on its platform, last week’s events raise a number of fundamental issues about online governance. Whether or not YouTube should adopt policies about hate speech and harassment on its platform, the fact is that it has adopted those guidelines and purports to take actions based on them.

But despite the presence of these policies, YouTube failed at every step of the Maza-Crowder debacle to communicate the basis of its actions clearly. Its initial conclusory tweet that Crowder’s videos did not violate YouTube’s policies were hard to square with the language of those policies. On their face, the videos Crowder published were “deliberately posted in order to humiliate” Maza and made “hurtful and negative personal comments” about him. The very fact that it took YouTube four days to respond suggests that this was not an easy call and, therefore, deserved more explanation. The partial flip-flopping regarding monetization and the selling of T-shirts was unpredictable and inexplicable. The company’s ultimate position, that Crowder’s videos did not violate its policies but “have harmed the broader community” is opaque and provides little guidance for users about what content YouTube will penalize.

Meanwhile, the limited explanations the company did give seem poorly thought through. Epithets framed as a “debate” or “comedic routine,” YouTube’s comments suggested, do not violate its policies. But this creates a standard so subjective as to seem unworkable in practice. YouTube says that its guidelines are about fostering the “trust” involved in keeping the YouTube community “fun and enjoyable for everyone.” It counsels people to take a purposive approach to interpreting them: “Don’t try to look for loopholes or try to lawyer your way around the guidelines—just understand them and try to respect the spirit in which they were created.” But despite its own advice not to “look for loopholes” its own guidelines, YouTube tied itself in knots, creating the appearance that the community guidelines are not so much rules as empty words that the company can interpret however it chooses.

These events showed a hard truth: Without government regulation, there is nothing requiring YouTube to set or abide by clear policies it holds out as the rules of its platform. The past few years have seen growing momentum behind calls for platforms to uphold the human rights of their users, including their rights not just to freedom of expression but also to due process. Many of the complaints about YouTube’s handling of the Maza-Crowder situation focused on the deficient processes and explanations that accompanied the company’s actions. The Verge’s Casey Newton called on YouTube to “have these arguments with us in public.” Former Facebook Chief Security Officer Alex Stamos said that YouTube needed “much more transparency in how these decisions are made. They need to document the thinking process, the tests they are using and the precedents they believe they are creating.”
Current company employees bemoaned the public trust that is lost when platforms fail to explain their decisions. The UN Special Rapporteur on freedom of expression suggested affected users deserved a more fully reasoned response than YouTube’s initial conclusory statement that Crowder’s videos did not violate its policies.

Even if it would be normatively desirable for content moderations to be transparent and accountable, there is no lever in the current legal landscape to enforce these calls. I have written before about proposed regulations in other countries that may change this. But currently, as Sarah Jeong put it in the New York Times, “YouTube is entitled to shoot entirely from the hip.” Nothing mandates YouTube to provide an avenue to productively channel or manage user grievance.

Laboratories of Online Governance

Justice Louis Brandeis famously praised federalism as allowing states to be laboratories of democracy, trying different approaches to rules to discover what works best. The governors of large online spaces now appear to be taking a similar approach. Although there remain superficial similarities between the large social media platforms, their approaches to dealing with the difficulties of content moderation look set to diverge in important ways. In response to many of the same criticisms YouTube faced in the past week, Facebook has announced it will create an oversight board, independent from the company, to hear appeals to its content moderation decisions with the goal of bringing exactly the kind of transparency and accountability to Facebook’s content moderation ecosystem that YouTube lacked.

Where YouTube’s decisions appeared motivated by commercial considerations, with observers speculating some accounts become “too big to fail,” users might have more confidence that well-designed independent oversight would be concerned first and foremost with the application of the publicly available rules. Where YouTube seemed ill-prepared to defend its actions, resorting to confusing tweets, an oversight board’s purpose would be to hear disputes and provide public reasoning for decisions. And these reasons would be focused on establishing workable precedents (rather than subjective standards such as whether a slur was done in the course of “debate”) to create a coherent body of platform law. With proper institutional design, the decision-makers would be less susceptible to public pressure, which might prevent disputes from descending into such painful online battles between opposing camps. All of this would engender greater public legitimacy in the platform’s policies and decisions.

Given these benefits, there was speculation last week that YouTube might also adopt a similar institutional mechanism. But the public legitimacy of creating independent oversight comes at the cost of allowing business reasons to govern those decisions. Ultimately, the Facebook oversight represents a bet: that the public legitimacy it might create by introducing a check and balance into its content moderation system is a good long-term investment to stem public controversy and create user buy-in for the platform’s rules. YouTube’s actions last week suggest a different gamble. The ad hoc, reactive and short-term thinking that seemed to pervade YouTube’s response indicates public legitimacy is not a governing concern for its decision-makers.

YouTube’s wager last week was not new, but until now the payoffs have seemed to go the other way. YouTube’s parent company, Google, did not even send a representative to Senate committee hearings with social platforms about foreign influence operations on social media, leaving Facebook and Twitter representatives to take most of the heat. When a video of House Speaker Nancy Pelosi, altered to make her appear drunk, started circulating on social media recently, Facebook decided not to remove the videos and rolled out a high-level executive to explain the decision on national television. Facebook was widely criticized for both its decision and the executive’s attempts to explain it. Meanwhile, YouTube quietly took down the video. The ultimate explanation YouTube gave—that the Pelosi video violated the company’s deceptive practices policy—was what Newton generously described as “a little slippery.” Yet YouTube largely escaped public scrutiny—only to face a firestorm of criticism around the Maza-Crowder issue weeks later.

In principle, however, YouTube’s lack of transparency and accountability regarding its decision to take down the Pelosi video should be as problematic as the Maza-Crowder decisions. Similarly, YouTube’s pro forma statement that it had decided to remove Alex Jones’s channel without explanation, right when a number of platforms were suddenly making the same decision, was, by the same standards, also procedurally illegitimate.

Among the theories animating the current push to break up the large social media platforms is the notion of “laboratories of online governance”—the idea that competition will allow greater experimentation and allow “healthier, less exploitative social media platforms” to emerge. This may be so. But YouTube’s actions over the past week are a case study in showing that, without regulatory guardrails, there is no guarantee that the platforms that emerge will focus on legitimizing the way they exercise their power.
On June 5, in the middle of the ongoing fallout from Maza’s complaints, YouTube released a new policy on hateful and supremacist content, announcing that it was expanding its rules to ban videos promoting Nazi ideology, Holocaust denial or Sandy Hook conspiracy theories. The platform stated in the announcement that “context matters,” so condemnation or analysis of hate could stay up. But in the rollout of the new rules, YouTube removed benign content including educational videos from a history teacher that included archival Nazi footage and an independent journalist’s videos documenting extremism.

At the end of a very long week for YouTube, many unanswered questions remained. Chief among them is what the value of YouTube’s new policy will be if the community doubts YouTube has either the capacity or the intention to actually enforce the policy as written.
How to Make Facebook’s ‘Supreme Court’ Work

The idea of a body that will decide what kind of content is allowed on the site is promising — but only if it’s done right.

By Kate Klonick and Thomas Kadri
Dr. Klonick and Mr. Kadri are lawyers.

Nov. 17, 2018

Facebook’s chief executive, Mark Zuckerberg, has announced a plan to create an independent body to make decisions about what kinds of content his site’s users will be allowed to post. It’s a development that he hinted at when he said on a podcast in April, “You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don’t work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world.”

It appears that this “Supreme Court” idea is now going to be carried out — and the independence, transparency, accountability and oversight Mr. Zuckerberg described in announcing on Thursday that such a body would be created within a year all sound promising. But details so far are scarce, and what this really means for free speech and fair process on the internet will depend on the answer to one key question: How much will the “Supreme Court of Facebook” be like the Supreme Court of the United States?

In theory, a court has at least three virtues. The first is due process: such a court can allow people to argue that mistakes have been made, and the court can then publicly explain its final decision. The second virtue is representation: the justices can represent different segments of society, bringing diverse perspectives and expertise to the difficult questions that they must answer. The third virtue is independence: while a legislature debates and passes laws, a court can be insulated from this political process when it interprets those laws and resolves competing legal claims.

There are good reasons to doubt whether even the United States Supreme Court is an institution that lives up to these ideals. But at the very least there are structures in place that aim to promote these values on the court. On Thursday, Mr. Zuckerberg admitted that Facebook is still figuring out how its panel will function. “Starting today,” he said, “we’re beginning a consultation period to address the hardest questions” about “how this will work in practice.” As Facebook — which operates in many ways like a government when it determines what kinds of speech are allowed — does so, it should pay attention to the government that already exists.

Questions abound about how this new tribunal will protect due process. For one, in trying to sniff out when mistakes have been made, what “record” of evidence will Facebook’s justices consider? In our legal system, the record is developed in the lower courts long before the Supreme Court hears the appeal, and the justices don’t get to do their own fact-finding to figure out what happened in the case before them.

But one of the greatest challenges posed by moderation of speech on Facebook is how heavily these decisions depend on context; the difference between a racist slur and a rap lyric, for example, might turn on the speaker’s identity, her motivations, her audience. These challenges become even more complex in a global context in which moderators must
account for different languages and slang; for different historical, cultural and political divides; and for different power structures — all of which might color the social meaning of the speech.

What does this mean for Facebook's new oversight council? In short, you need real-world evidence to get real-world context. The Silicon Valley “justices” adjudicating questions of online speech will need fact-finding powers that we don't give to their counterparts in Washington. If all that they review is the same lifeless screenshot seen by Facebook's own moderators, it's tough to see how they'll be in a much better position to correct mistakes. That may give Facebook's users more process, but is that really the process they are due?

Then there’s the issue of representation: Who will Facebook's justices be, and how will they be chosen? As we saw with the recent confirmation battle over Brett Kavanaugh, the identity of our judges matters a great deal. We subject nominees to extreme scrutiny, and the Constitution splits the authority to appoint justices between two branches of government to offer at least some check on that awesome power.

Given the diversity of Facebook's community, the body should be international, represent multiple stakeholders, and include voices from groups that are targets of hate speech and harassment on social media. But that still leaves the problem of “who.” A 100-person panel might represent Facebook's diverse community better than the nine-justice Supreme Court represents America, but such a vast institution would surely struggle to deliberate and decide on the questions before it. Still, Facebook's council can't hope to gain legitimacy if it doesn't represent a broad array of viewpoints and perspectives.

Finally, in considering the issue of independence, Facebook can learn a few things from our nation's highest court. The Supreme Court is part of something greater — it is the third branch in a system that creates a separation of powers. Congress makes laws, the court interprets them, and the executive enforces them. Here, Facebook would, in a sense, still be playing the role of both Congress and the executive. But the court's role in checking congressional and executive actions is crucial, and it works only because the court is independent from the other branches. The justices have broad discretion in deciding which cases to hear, and justices are appointed for life in an attempt to shield them from the pressures of outside politics.

The new council imagined by Facebook feels like an attempt to create this type of independent body. Indeed, Mr. Zuckerberg says that Facebook is creating this organization to “prevent the concentration of too much decision-making within our teams” and to “provide assurance that these decisions are made in the best interests of our community and not for commercial reasons.” But it is also in the best interests of Facebook that such a tribunal would be a convenient scapegoat for contentious decisions: “Don't like how we dealt with the takedown of the Alex Jones pages? Don't blame us! It was the Council!”

For Facebook's new appeals body to be more than an empty gesture, it must be independent from its creators. It's obvious that such an oversight group must be independently funded — Facebook can't control the purse strings — but that's not enough. There are two other critical features that can give a court independence from other governmental branches. One is the court's ability to decide which appeals it hears, and Facebook's tribunal should also have broad discretion to pick its “cases.”

But the second and perhaps much more powerful part comes from a court's adherence to a constitution. A constitution plays
many roles, but one of them is its ability to be stalwart on values in the face of societal change: Constitutions are difficult to amend and difficult to reinterpret. Whatever direction the wind blows, or whatever Sirens may call, a constitution will be the proverbial binding that tethers Odysseus to the mast and ensures that the ship continues to sail true.

With that in mind, Facebook should consider — especially if it continues to act as a type of governing body — adopting something like a constitution that is harder to amend that its ever-shifting content-moderation rules, which it could alter mercurially to get around decisions issued by its court that it doesn't like.

The idea of a Supreme Court of Facebook is promising in theory. But how all this will function ultimately rests on choices that Mr. Zuckerberg has yet to make. All we can do is hope that he chooses wisely — and therein lies the perilous relationship we have with Facebook.

Kate Klonick is an assistant professor at St. John's Law School and the author of “The New Governors: The People, Rules, and Processes Governing Online Speech.” Thomas Kadri is a resident fellow at Yale Law School's Information Society Project and a doctoral candidate at Yale Law School.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.
THE NEW GOVERNORS: THE PEOPLE, RULES, AND PROCESSES GOVERNING ONLINE SPEECH

Kate Klonick

INTRODUCTION .......................................................................................................................... 1599

I. SECTION 230, THE FIRST AMENDMENT, AND THE BEGINNINGS OF INTERMEDIARY SELF-REGULATION .............................................................................. 1603
   A. History and Development of § 230 .................................................................................. 1604
   B. First Amendment Implications ..................................................................................... 1609
   C. Internet Pessimists, Optimists, and Realists ................................................................. 1613

II. WHY GOVERN WELL? THE ROLE OF FREE SPEECH NORMS, CORPORATE CULTURE, AND ECONOMIC INCENTIVES IN THE DEVELOPMENT OF CONTENT MODERATION...................................................................................................................... 1616
   A. Platforms' Baseline in Free Speech .............................................................................. 1618
      1. Free Speech Norms ................................................................................................... 1618
      2. Government Request and Collateral Censorship Concerns .................................. 1622
   B. Why Moderate At All? .................................................................................................... 1625
      1. Corporate Responsibility and Identity ....................................................................... 1626
      2. Economic Reasons ..................................................................................................... 1627

III. HOW ARE PLATFORMS GOVERNING? THE RULES, PROCESS, AND REVISION OF CONTENT-MODERATION SYSTEMS .............................................................................. 1630
   A. Development of Moderation: From Standards to Rules ............................................ 1631
   B. How the Rules Are Enforced: Trained Human Decisionmaking.............................. 1635
      1. Ex Ante Content Moderation ....................................................................................... 1636
      2. Ex Post Proactive Manual Content Moderation ....................................................... 1638
      3. Ex Post Reactive Manual Content Moderation ....................................................... 1638
      4. Decisions, Escalations, and Appeals ......................................................................... 1647
   C. System Revision and the Pluralistic System of Influence ............................................ 1649
      1. Government Requests ................................................................................................. 1650
      2. Media Coverage ......................................................................................................... 1652
      3. Third-Party Influences ............................................................................................... 1655
      4. Change Through Process ........................................................................................... 1657
   D. Within Categories of the First Amendment ................................................................ 1658

IV. THE NEW GOVERNORS .................................................................................................... 1662
   A. Equal Access ................................................................................................................ 1665
   B. Accountability .............................................................................................................. 1666

CONCLUSION ............................................................................................................................ 1669
Private online platforms have an increasingly essential role in free speech and participation in democratic culture. But while it might appear that any internet user can publish freely and instantly online, many platforms actively curate the content posted by their users. How and why these platforms operate to moderate speech is largely opaque.

This Article provides the first analysis of what these platforms are actually doing to moderate online speech under a regulatory and First Amendment framework. Drawing from original interviews, archived materials, and internal documents, this Article describes how three major online platforms — Facebook, Twitter, and YouTube — moderate content and situates their moderation systems into a broader discussion of online governance and the evolution of free expression values in the private sphere. It reveals that private content-moderation systems curate user content with an eye to American free speech norms, corporate responsibility, and the economic necessity of creating an environment that reflects the expectations of their users. In order to accomplish this, platforms have developed a detailed system rooted in the American legal system with regularly revised rules, trained human decisionmaking, and reliance on a system of external influence.

This Article argues that to best understand online speech, we must abandon traditional doctrinal and regulatory analogies and understand these private content platforms as systems of governance. These platforms are now responsible for shaping and allowing participation in our new digital and democratic culture, yet they have little direct accountability to their users. Future intervention, if any, must take into account how and why these platforms regulate online speech in order to strike a balance between preserving the democratizing forces of the internet and protecting the generative power of our New Governors.

INTRODUCTION

In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we’re really setting policies.

— Mark Zuckerberg

†

† Ph.D. in Law Candidate, Yale University, and Resident Fellow at the Information Society at Yale Law School. Research for this project was made possible with the generous support of the Oscar M. Ruebhausen Fund. The author is grateful to Jack Balkin, Molly Brady, Kiel Brennan-Marquez, Peter Byrne, Adrian Chen, Bryan Choi, Danielle Keats Citron, Rebecca Crootof, Evelyn Frazee, Tarleton Gillespie, Eric Goldman, James Grimmelmann, Brad Greenberg, Alexandra Gutierrez, Woody Hartzog, David Hoffman, Gus Hurwitz, Thomas Kadri, Margot Kaminski, Alyssa King, Jonathan Manes, Toni Massaro, Christina Mulligan, Frank Pasquale, Robert Post, Sabeel Rahman, Jeff Rosen, Andrew Selbst, Jon Shea, Rebecca Tushnet, and Tom Tyler for helpful thoughts and comments on earlier versions of this Article. A special thank you to Rory Van Loo, whose own paper workshop inadvertently inspired me to pursue this topic. Elizabeth Goldberg and Deborah Won provided invaluable and brilliant work as research assistants.

In the summer of 2016, two historic events occurred almost simultaneously: a bystander captured a video of the police shooting of Alton Sterling on his cell phone, and another recorded the aftermath of the police shooting of Philando Castile and streamed the footage via Facebook Live. Following the deaths of Sterling and Castile, Facebook founder and CEO Mark Zuckerberg stated that the ability to instantly post a video like the one of Castile dying “reminds us why coming together to build a more open and connected world is so important.” President Barack Obama issued a statement saying the shootings were “symptomatic of the broader challenges within our criminal justice system,” and the Department of Justice opened an investigation into Sterling’s shooting and announced that it would monitor the Castile investigation. Multiple protests took place across the country. The impact of these videos is an incredible example of how online platforms are now essential to participation in democratic culture. But it almost never happened.

Initially lost in the voluminous media coverage of these events was a critical fact: as the video of Castile was streaming, it suddenly disappeared from Facebook. A few hours later, the footage reappeared, this time with a label affixed warning of graphic content. In official state-
ments, Facebook blamed the takedown on a “technical glitch” but provided no further details. This is not entirely surprising. Though it might appear that any internet user can publish freely and instantly online, many content-publication platforms actively moderate the content posted by their users. Yet despite the essential nature of these platforms to modern free speech and democratic culture, very little is known about how or why these companies curate user content.

In response to calls for transparency, this Article examines precisely what these private platforms are actually doing to moderate user-generated content and why they are doing so. It argues that these platforms are best thought of as self-regulating private entities, governing

---

11 I use the terms “moderate,” “curate,” and sometimes “regulate” to describe the behavior of these private platforms in both keeping up and taking down user-generated content. I use these terms rather than using the term “censor,” which evokes the ideas of only removal of material and various practices of culturally expressive discipline or control. See generally Robert C. Post, Project Report: Censorship and Silencing, 51 Bull. Am. Acad. Arts & Sci. 32, 32 (1998). Where I do use “regulate,” I do so in a more colloquial sense and not the way in which Professor Jack Balkin uses the term “speech regulation,” which concerns government regulation of speech or government cooperation, coercion, or partnership with private entities to reflect government ends. See Jack M. Balkin, Old-School/New-School Speech Regulation, 127 Harv. L. Rev. 2296, 2299 (2014) (also explaining that the phrase “collateral censorship” is a term of art exempted from this taxonomy).
13 Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017) (holding that a state statute barring registered sex offenders from using online social media platforms was unconstitutional under the First Amendment). In his majority opinion, Justice Kennedy wrote that “[w]hile in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyberspace — the ‘vast democratic forums of the Internet’ in general, and social media in particular.” Id. at 1735 (citation omitted) (quoting Reno v. ACLU, 521 U.S. 844, 868 (1997)).
14 See, e.g., Marvin Ammori, The “New” New York Times: Free Speech Lawyering in the Age of Google and Twitter, 127 Harv. L. Rev. 2259, 2273–76 (2014); Marjorie Heins, The Brave New World of Social Media Censorship, 127 Harv. L. Rev. F. 325, 362 (2014) (describing Facebook’s internal appeals process as “mysterious at best” and noting, about their internal policies, that “[t]he details of these rules . . . we do not know” and that the censorship “process in the private world of social media is secret”).
speech within the coverage of the First Amendment\textsuperscript{16} by reflecting the
democratic culture and norms of their users.\textsuperscript{17}

Part I surveys the regulatory and constitutional protections that have
resulted in these private infrastructures. The ability of private platforms
to moderate content comes from § 230 of the Communications Decency
Act\textsuperscript{18} (CDA), which gives online intermediaries broad immunity from
liability for user-generated content posted on their sites.\textsuperscript{19} The purpose
of this grant of immunity was both to encourage platforms to be “Good
Samaritans” and take an active role in removing offensive content, and
also to avoid free speech problems of collateral censorship.\textsuperscript{20} Beyond
§ 230, courts have struggled with how to conceptualize online platforms
within First Amendment doctrine: as state actors, as broadcasters, or as
editors. Additionally, scholars have moved between optimistic and pes-
simistic views of platforms and have long debated how — or whether —
to constrain them.

To this legal framework and scholarly debate, this Article applies
new evidence. Part II looks at why platforms moderate so intricately
given the broad immunity of § 230. Through interviews with former
platform architects and archived materials, this Article argues that plat-
forms moderate content because of a foundation in American free speech
norms, corporate responsibility, and the economic necessity of creating
an environment that reflects the expectations of their users. Thus, plat-
forms are motivated to moderate by both of § 230’s purposes: fostering
Good Samaritan platforms and promoting free speech.

Part III looks at how platforms are moderating user-generated con-
tent and whether that understanding can fit into an existing First
Amendment framework. Through internal documents, archived mate-
rials, interviews with platform executives, and conversations with con-
tent moderators, this Article shows that platforms have developed a sys-
tem that has marked similarities to legal or governance systems. This
includes the creation of a detailed list of rules, trained human deci-
sionmaking to apply those rules, and reliance on a system of external
influence to update and amend those rules. With these facts, this Article

\textsuperscript{16} See generally Balkin, supra note 11; Frederick Schauer, The Boundaries of the First

\textsuperscript{17} See generally ROBERT C. ELICKSON, ORDER WITHOUT LAW (1991); ELINOR OSTROM,
CRAFTING INSTITUTIONS FOR SELF-GOVERNING IRRIGATION SYSTEMS (1992); Balkin, su-
pra note 7; J.M. Balkin, Populism and Progressivism as Constitutional Categories, 104 YALE L.J.
FREE SPEECH (1993), and defining democratic culture as popular participation in culture); Robert
C. Ellickson, Of Coase and Cattle: Dispute Resolution Among Neighbors in Shasta County, 38


\textsuperscript{19} Id.

\textsuperscript{20} See Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (noting that the purposes
of intermediary immunity in § 230 were not only to incentivize platforms to remove indecent content
but also to protect the free speech of platform users).
argues that analogy under purely First Amendment doctrine should be largely abandoned.

Instead, platforms should be thought of as operating as the New Governors of online speech. These New Governors are part of a new triadic model of speech that sits between the state and speakers-publishers. They are private, self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free speech expectations of their users. Part IV explains how this conceptualization of online platforms as governance fits into scholarly concerns over the future of digital speech and democratic culture. It argues that the biggest threat this private system of governance poses to democratic culture is the loss of a fair opportunity to participate, which is compounded by the system’s lack of direct accountability to its users. The first solution to this problem should not come from changes to § 230 or new interpretations of the First Amendment, but rather from simple changes to the architecture and governance systems put in place by these platforms. If this fails and regulation is needed, it should be designed to strike a balance between preserving the democratizing forces of the internet and protecting the generative power of our New Governors, with a full and accurate understanding of how and why these platforms operate, as presented here. It is only through accurately understanding the infrastructures and motivations of our New Governors that we can ensure that the free speech rights essential to our democratic culture remain protected.

I. SECTION 230, THE FIRST AMENDMENT, AND THE BEGINNINGS OF INTERMEDIARY SELF-REGULATION

Before the internet, the most significant constraint on the impact and power of speech was the publisher.21 The internet ended the speaker’s reliance on the publisher by allowing the speaker to reach his or her audience directly.22 Over the last fifteen years, three American companies — YouTube, Facebook, and Twitter — have established themselves as dominant platforms in global content sharing.23 These platforms are both the architecture for publishing new speech and the architects of the

21 LAWRENCE LESSIG, CODE 2.0, at 19 (2006).
22 Id., Balkin, supra note 11, at 2306–10.
institutional design that governs it.24 This private architecture is the “central battleground over free speech in the digital era.”25

A. History and Development of § 230

In order to understand the private governance systems used by platforms to regulate user content, it is necessary to start with the legal foundations and history that allowed for such a system to develop. The broad freedom of internet intermediaries26 to shape online expression is based in § 230 of the CDA, which immunizes providers of “interactive computer services” from liability arising from user-generated content.27 Sometimes called “the law that matters most for speech on the Web,” the existence of § 230 and its interpretation by courts have been essential to the development of the internet as we know it today.28

Central to understanding the importance of § 230 are two cases decided before its existence, which suggested that intermediaries would be liable for defamation posted on their sites if they actively exercised any editorial discretion over offensive speech.29 The first, Cubby, Inc. v. CompuServe, Inc.,30 involved the publication of libel on CompuServe forums.31 The court found CompuServe could not be held liable for the defamatory content in part because the intermediary did not review any of the content posted to the forum.32 The Cubby court reasoned that CompuServe’s practice of not actively reviewing content on its site made it more like a distributor of content, and not a publisher.33 In determining communication tort liability, this distinction is important because while publishers and speakers of content can be held liable, distributors are generally not liable unless they knew or should have known of the

24 LESSIG, supra note 21, at 2–10 (describing the internet as architecture).
25 Balkin, supra note 11, at 2296.
26 Internet intermediaries are broadly defined as actors in every part of the internet “stack.” See JAMES GRIMMELMANN, INTERNET LAW 31 (2016). These include internet service providers, hosting providers, servers, websites, social networks, search engines, and so forth. See id. at 31–32. Within this array, I use “platforms” to refer specifically to internet websites or apps that publish user content — these include Facebook, YouTube, and Twitter.
27 47 U.S.C. § 230(c)(2) (2012); see also Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (blocking claims against AOL under § 230 because AOL was only the publisher, and not the creator, of the tortious content).
28 Emily Bazelon, How to Unmask the Internet’s Vilest Characters, N.Y. TIMES MAG. (Apr. 22, 2011), http://nyti.ms/2C3oZLq [https://perma.cc/55A3-6FAN].
31 Id. at 138; Ardia, supra note 29, at 406–07. CompuServe did not dispute that the statements were defamatory. Cubby, 776 F. Supp. at 138.
32 Cubby, 776 F. Supp. at 140.
33 Id. at 139–41.
defamation.\textsuperscript{34} Though distributor-publisher distinctions were an established analogy in tort liability, the difficulty of using this model for online intermediaries quickly became apparent. Four years after Cubby, in \textit{Stratton Oakmont, Inc. v. Prodigy Services Co.},\textsuperscript{35} a court found that the intermediary Prodigy was liable as a publisher for all posts made on its site because it actively deleted some forum postings.\textsuperscript{46} To many, Prodigy’s actions seemed indistinguishable from those that had rendered CompuServe a mere distributor in Cubby, but the court found Prodigy’s use of automatic software and guidelines for posting were a “conscious choice, to gain the benefits of editorial control.”\textsuperscript{37} Read together, the cases seemed to expose intermediaries to a wide and unpredictable range of tort liability if they exercised any editorial discretion over content posted on their sites. Accordingly, the cases created a strong disincentive for online intermediaries to expand business or moderate offensive content and threatened the developing landscape of the internet.

Thankfully, the developing landscape of the internet was an active agenda item for Congress when the \textit{Stratton Oakmont} decision came down. Earlier that year, Senator James Exon had introduced the CDA, which aimed to regulate obscenity online by making it illegal to knowingly send or show minors indecent online content.\textsuperscript{38} Reacting to the concerns created by \textit{Stratton Oakmont}, Representatives Chris Cox and Ron Wyden introduced an amendment to the CDA that would become § 230.\textsuperscript{39} The Act, with the Cox-Wyden amendment, passed and was signed into law in February 1996.\textsuperscript{40} In its final form, § 230(c) stated that “no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider”\textsuperscript{41} in order to incentivize and protect intermediaries’ Good Samaritan blocking of offensive material.\textsuperscript{42} Though, just a little over a year later, the Supreme Court in \textit{Reno v. ACLU}\textsuperscript{43} struck down the bulk of the anti-indecency sections of the CDA, § 230 survived.\textsuperscript{44}

\begin{thebibliography}{1}
\bibitem{34} \textit{RESTATEMENT (SECOND) OF TORTS} § 581(1) (AM. LAW INST. 1977).
\bibitem{36} \textit{Id.} at *4.
\bibitem{37} \textit{Id.} at *5.
\bibitem{42} 141 CONG. REC. H8469–70 (statement of Rep. Cox).
\bibitem{43} 521 U.S. 844 (1997).
\bibitem{44} \textit{Id.} at 885.
\end{thebibliography}
It soon became clear that § 230 would do more than just survive. A few months after Reno, the Fourth Circuit established a foundational and expansive interpretation of § 230 in Zeran v. America Online, Inc.45 Plaintiff Zeran sought to hold AOL liable for defamatory statements posted on an AOL message board by a third party.46 Zeran argued that AOL had a duty to remove the posting, post notice of the removed post’s falsity, and screen future defamatory material.47 The court disagreed. Instead, it found AOL immune under § 230 and held that the section precluded not only strict liability for publishers but also intermediary liability for distributors such as website operators.48 This holding also extinguished notice liability for online intermediaries.49

While the holdings in Zeran were broad and sometimes controversial,50 it was the court’s analysis as to the purposes and scope of § 230 that truly shaped the doctrine. In granting AOL the affirmative defense of immunity under § 230, the court recognized the Good Samaritan provision’s purpose of encouraging “service providers to self-regulate the dissemination of offensive material over their services.”51 But the court did not consider § 230 merely a congressional response to Stratton Oakmont. Instead, the court looked to the plain text of § 230(c) granting statutory immunity to online intermediaries and drew new purpose beyond the Good Samaritan provision and found that intent “not difficult to discern”:

Congress recognized the threat that tort-based lawsuits pose to freedom of speech in the new and burgeoning Internet medium. The imposition of tort liability on service providers for the communications of others represented, for Congress, simply another form of intrusive government regulation of speech.52

Thus, while the court reasoned that § 230 lifted the “specter of tort liability” that might “deter service providers from blocking and screening offensive material,” it found it was also Congress’s design to immunize intermediaries from any requirement to do so.53 Drawing on these free speech concerns, the court reasoned that the same “specter of tort liability” that discouraged intermediaries from policing harmful content also threatened “an area of such prolific speech” with “an obvious

45 129 F.3d 327 (4th Cir. 1997).
46 Id. at 328.
47 Id. at 330.
48 Id. at 332.
49 Id. at 333.
51 Zeran, 129 F.3d at 331.
52 Id. at 330 (emphases added).
53 Id. at 331.
chilling effect.”54 “Faced with potential liability for each message re-published by their services, interactive computer service providers might choose to severely restrict the number and type of messages posted.”55 In response to the question raised by Zeran of subjecting publishers like AOL to notice-based liability, the court again cited its free speech concerns but also recognized the practical realities of distributors: “Each notification would require a careful yet rapid investigation of the circumstances surrounding the posted information, a legal judgment concerning the information’s defamatory character, and an on-the-spot editorial decision whether to risk liability by allowing the continued publication of that information.”56

The sheer volume of content to be policed by intermediaries, and their almost certain liability should they be notified and still publish, would lead to either haphazard takedowns at best, or widespread removal at worst. “Thus, like strict liability, liability upon notice has a chilling effect on the freedom of Internet speech.”57

Zeran is a seminal decision in internet law not only because it gave broad immunity to online intermediaries58 but also because of its analysis of the purposes of § 230. The court recognized two distinct congressional purposes for granting immunity under § 230: (1) as a Good Samaritan provision written to overturn Stratton Oakmont and “to encourage interactive computer services and users of such services to self-policing the Internet for obscenity and other offensive material,”59 and (2)

54 Id.
55 Id. The quote continues: “Congress considered the weight of the speech interests implicated and chose to immunize service providers to avoid any such restrictive effect.” Id.
56 Id. at 333.
57 Id. Though this free speech purpose might not have been in the plain text of § 230, the Zeran court did not invent it. See Cannon, supra note 38, at 88–91 (discussing the legislative history indicating that Congress debated the “contest between censorship and democratic discourse,” id. at 88).
as a free speech protection for users meant “to encourage the unfettered and unregulated development of free speech on the Internet, and to promote the development of e-commerce.”

Though the exact term is not stated in the text of Zeran, the court’s concern over service providers’ “natural incentive simply to remove messages upon notification, whether the contents were defamatory or not,” reflects apprehension of collateral censorship. Collateral censorship occurs when one private party, like Facebook, has the power to control speech by another private party, like a Facebook user. Thus, if the government threatens to hold Facebook liable based on what its user says, and Facebook accordingly censors its user’s speech to avoid liability, you have collateral censorship. The court in Zeran recognized this concern for the free speech rights of users and counted it among the reasons for creating immunity for platforms under §230.

But while the dual purposes of §230 call for the same solution — intermediary immunity — they create a paradox in the applications of §230. If §230 can be characterized as both government-created immunity to (1) encourage platforms to remove certain kinds of content, and (2) avoid the haphazard removal of certain content and the perils of collateral censorship to users, which interests do we want to prioritize? That of the platforms to moderate their content or that of users’ free speech?

In the last few years, courts have grappled with precisely this dilemma and occasionally broken with the expansive interpretation of the Good Samaritan provision to find a lack of §230 immunity. For instance, in two recent district court cases in northern California, the court

and Barton); then citing Zeran, 129 F.3d at 331; and then citing Blumenthal v. Drudge, 992 F. Supp. 44, 52 (D.D.C. 1998)).

60 Id. at 1027–28 (first citing §230(b) (policy objectives include “(1) to promote the continued development of the Internet and other interactive computer services and other interactive media; (2) to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation”); then citing Zeran, 129 F.3d at 330).

61 Zeran, 129 F.3d at 333. The court also specifically cited worry about potential abuse between users. “Whenever one was displeased with the speech of another party conducted [online], the offended party could simply ‘notify’ the relevant service provider, claiming the information to be legally defamatory.” Id.; see also Christina Mulligan, Technological Intermediaries and Freedom of the Press, 66 SMU L. REV. 157, 173 (2013); Felix T. Wu, Collateral Censorship and the Limits of Intermediary Immunity, 87 NOTRE DAME L. REV. 293, 317–18 (2011).

62 The term “collateral censorship” was coined by Professor Michael Meyerson. Michael I. Meyerson, Authors, Editors, and Uncommon Carriers: Identifying the “Speaker” Within the New Media, 71 NOTRE DAME L. REV. 79, 118 (1995).


64 For a comprehensive cataloging of §230 cases with context and commentary, see Eric Goldman, Ten Worst Section 230 Rulings of 2016 (Plus the Five Best), TECH. & MARKETING L.
rejected motions to dismiss for failure to state a claim under § 230 on the basis of plaintiffs’ allegations that Google acted in bad faith. At the same time, other courts have made powerful decisions in favor of broad § 230 immunity and publishers’ rights to moderate content. Notably, in Doe v. Backpage.com, the First Circuit expressly held that § 230 protects the choices of websites as speakers and publishers, stating: “Congress did not sound an uncertain trumpet when it enacted the CDA, and it chose to grant broad protections to internet publishers. Showing that a website operates through a meretricious business model is not enough to strip away those protections.” The continued confusion about § 230’s interpretation — as seen in current courts’ split on the importance of a business’s motivations for content moderation — demonstrates that the stakes around such questions have only grown since the foundational decision in Zeran.

B. First Amendment Implications

The debate over how to balance the right of intermediaries to curate a platform while simultaneously protecting user speech under the First Amendment is ongoing for courts and scholars. Depending on the type of intermediary involved, courts have analogized platforms to established doctrinal areas in First Amendment law — company towns, broadcasters, editors — and the rights and obligations of a platform shift depending on which analogy is applied.

The first of these analogies reasons that platforms are acting like the state, so the First Amendment directly constrains them. While courts have established that only state action creates affirmative obligations under the First Amendment, determining exactly when a private party’s behavior constitutes state action is a more difficult question. The Supreme Court foundationally addressed this distinction between private and state actors for First Amendment purposes in Marsh v. Alabama. In Marsh, a Jehovah’s Witness was arrested for criminal trespass for distributing literature on the sidewalk of a company town


66 817 F.3d 12 (1st Cir. 2016).

67 Id. at 29.


wholly owned by a corporation. The Court found that “[e]xcept for ownership by a private corporation, this town] has all the characteristics of any other American town.” Accordingly, the Court held the town was functionally equivalent to a state actor and obligated to guarantee First Amendment rights.

In the years since Marsh, the Court has continued to explore the “public function” circumstances necessary for private property to be treated as public. Many of these cases have arisen in the context of shopping malls, where the Court has struggled to establish consistent reasoning on when a private individual’s First Amendment rights trump the rights of the owner of a private forum. The most expansive of these was Amalgamated Food Employees Union Local 590 v. Logan Valley Plaza, Inc., which held a shopping mall to be the equivalent of the company town in Marsh and therefore allowed picketers to protest there. In overruling Logan Valley in Hudgens v. NLRB, the Court revised its assessment of a shopping mall as a public square and stated that a business does not qualify as performing a public function merely because it is open to the public. Instead, in order to qualify as performing a public function, a business must be actually doing a job normally done by the government, as was the case with the company town in Marsh.

For a long time, the claim that online intermediaries are state actors or perform a public function and, thus, are subject to providing free speech guarantees, was a losing one. In establishing platforms as non-state actors, courts distinguished the facts in Marsh and its progeny, stating that intermediaries providing services like email, hosting, or search engines do not rise to the level of “performing any municipal power or essential public service and, therefore, do[ ] not stand in the

70 Id. at 502–03.
71 Id. at 502.
72 Id. at 508–09.
73 See, e.g., Amalgamated Food Emps. Union Local 590 v. Logan Valley Plaza, Inc., 391 U.S. 308, 318 (1968) (equating a private shopping center to a business district and affirming the right to picket in it), narrowed by Lloyd Corp. v. Tanner, 407 U.S. 551, 563–64 (1972) (holding speech in a mall is not constitutionally protected unless there are no other means of communication), overruled by Hudgens, 424 U.S. at 518. The California Supreme Court granted more expansive free speech guarantees than those provided by the First Amendment in Fashion Valley Mall, LLC v. NLRB, 172 P.3d 742, 749 (Cal. 2007), and Robins v. PruneYard Shopping Center, 592 P.2d 341, 344, 347 (Cal. 1979). See also Developments in the Law — State Action and the Public/Private Distinction, 123 HARV. L. REV. 1248, 1303–07 (2010).
74 391 U.S. 308.
75 Id. at 318.
76 424 U.S. 507.
77 Id. at 519 (quoting Lloyd Corp., 407 U.S. at 568–69).
78 Id.
While these cases have not been explicitly overturned, the Court’s recent ruling in *Packingham v. North Carolina* might breathe new life into the application of state action doctrine to internet platforms.

In *Packingham*, the Court struck down a North Carolina statute barring registered sex offenders from platforms like Facebook and Twitter. In his opinion for the court, Justice Kennedy reasoned that foreclosing “access to social media altogether is to prevent the user from engaging in the legitimate exercise of First Amendment rights.” Describing such services as a “modern public square,” Justice Kennedy also acknowledged their essential nature to speech, calling them “perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard.” Though the decision is limited in that it applies only to total exclusion, the sweeping language makes access to private online platforms a First Amendment right, leaving open the questions of how robust that access must be or where in the internet pipeline a choke point must lie in order to abridge a First Amendment right. Future litigation might use *Packingham’s* acknowledgment of a First Amendment right to social media access as a new basis to argue that these platforms perform quasi-municipal functions.

Separate from the issue of state action, *Packingham’s* acknowledgment of platforms as private forums that significantly affect the expressive conduct of other private parties implicates other areas of regulation that are consistent with the First Amendment. This can be seen in the doctrine around other types of speech conduits, like radio and television broadcasters. In such cases, the Court has upheld regulation of radio broadcasting, despite the broadcast station’s claims that the regulation unconstitutionally infringed on its editorial judgment and speech. A public right to “suitable access” to ideas and a scarce radio spectrum justified the agency rule that required broadcasters to present public

---

79 Cyber Promotions, Inc. v. Am. Online, Inc., 948 F. Supp. 436, 442 (E.D. Pa. 1996) (distinguishing AOL’s email service from the kind of “municipal powers or public services” provided by a private company town that made it liable as a state actor in *Marsh*); see also Green v. Am. Online, 318 F.3d 465, 472 (3d Cir. 2003) (holding that, as a private company and not a state actor, AOL is not subject to constitutional free speech requirements); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 631 (D. Del. 2007) (finding that for the purposes of constitutional free speech guarantees, Google, Yahoo, and Microsoft are private companies, even though they work with state actors like public universities).

80 137 S. Ct. 1730 (2017).

81 Id. at 1733, 1738.

82 Id. at 1737.

83 Id.

issues and give each side of those issues fair coverage.\textsuperscript{85} In the years following, the Court has limited this holding,\textsuperscript{86} while also extending it to the realm of broadcast television in\textit{Turner Broadcasting System, Inc. v. FCC}.\textsuperscript{87}

The question of whether internet intermediaries would fall in the same category as radio or broadcast television was addressed by the Court in\textit{Reno}. The Court found that the elements that justify television and radio regulation — those mediums’ “invasive” nature, history of extensive regulation, and the scarcity of frequencies — “are not present in cyberspace” and explicitly exempted the internet from the doctrine established in\textit{Red Lion Broadcasting Co. v. FCC}\textsuperscript{88} and \textit{Turner}.\textsuperscript{89} While it is unclear how the Court would draw the line between the internet functions of concern in\textit{Reno} and the growth of social media platforms,\textit{Packingham}’s emphasis on the right to platform access might revive the concerns over scarcity raised by these cases.

The final First Amendment analogy relevant to online speech reasons that platforms themselves exercise an important expressive role in the world, so the First Amendment actively protects them from state interference. This draws on the doctrine giving special First Amendment protections to newspapers under\textit{Miami Herald Publishing Co. v. Tornillo}.\textsuperscript{90} There, in a unanimous decision, the Court found a Florida statute that gave political candidates a “right to reply” in local newspapers unconstitutional under the Free Press Clause of the First Amendment.\textsuperscript{91} Though the “right to reply” legislation was akin to FCC fairness regulations upheld in\textit{Red Lion}, the\textit{Tornillo} Court found the statute unconstitutional.\textsuperscript{92} The Court reasoned that the statute was an

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{85} Id. at 400–01 (“In view of the scarcity of broadcast frequencies, the Government’s role in allocating those frequencies, and the legitimate claims of those unable without governmental assistance to gain access to those frequencies for expression of their views, we hold the regulations and ruling at issue here are both authorized by statute and constitutional.”).
\item \textsuperscript{86} See, e.g., FCC v. League of Women Voters, 468 U.S. 364, 402 (1984) (holding publicly funded broadcasters have First Amendment protections to editorialize); FCC v. Pacifica Found., 438 U.S. 726, 741 n.17 (1978) (stating “it is well settled that the First Amendment has a special meaning in the broadcasting context” and citing\textit{Red Lion}); Columbia Broad. Sys., Inc. v. Democratic Nat’l Comm., 412 U.S. 94, 120–21 (1973) (holding broadcasters are not under an obligation to sell advertising time to a political party).
\item \textsuperscript{87} Turner Broad. Sys., Inc. v. FCC (\textit{Turner II}), 520 U.S. 180, 185 (1997); Turner Broad. Sys., Inc. v. FCC (\textit{Turner I}), 512 U.S. 622, 638–39 (1994). In these cases the Court dealt with FCC “must carry” regulations imposed on cable television companies. In\textit{Turner I}, the Court determined that cable television companies were indeed First Amendment speakers, 512 U.S. at 656, but in\textit{Turner II}, it held that the “must carry” provisions of the FCC did not violate those rights, 520 U.S. at 224–25.
\item \textsuperscript{88} 395 U.S. 367.
\item \textsuperscript{89} Reno v. ACLU, 521 U.S. 844, 868–70 (1997).
\item \textsuperscript{90} 418 U.S. 241 (1974).
\item \textsuperscript{91} Id. at 247, 258.
\item \textsuperscript{92} Id. at 258.
\end{itemize}
\end{footnotesize}
“intrusion into the function of editors” and that “press responsibility is not mandated by the Constitution and . . . cannot be legislated.” As internet intermediaries have become more and more vital to speech, First Amendment advocates have urged courts to apply the holding in Tornillo to platforms, granting them their own speech rights. The Court’s new definition in Packingham of online speech platforms as forums, however, might threaten the viability of arguments that these companies have their own First Amendment rights as speakers.

C. Internet Pessimists, Optimists, and Realists

As have the courts, scholars have struggled with the question of how to balance users’ First Amendment right to speech against intermediaries’ right to curate platforms. Many look to platforms as a new market for speech and ideas. In the early days of the internet, Professor Jack Balkin could have been considered an internet optimist. He saw the internet and its wealth of publishing tools, which enable widespread digital speech, as enhancing the “possibility of democratic culture.” More recently, he has recognized that private control of these tools poses threats to free speech and democracy. Professor Yochai Benkler could also have been considered an optimist, though a more cautious one. He has posited looking at the internet as enabling new methods of information production, as well as a move from traditional industrial-dominated markets to more collaborative peer production. Professor Lawrence Lessig acknowledges that while the internet creates exciting new means to regulate through code, he is concerned about corporations and platforms having great unchecked power to regulate the internet and all interactions that fall under § 230 immunity. Professors James Boyle, Jack Goldsmith, and Tim Wu have had similar concerns about

93 Id.
94 Id. at 256.
96 Balkin, supra note 7, at 45–46.
97 See Balkin, supra note 11, at 2300–01.
the state coopting private online intermediaries for enforcement. 100 Professor David Post has argued that the market will resolve corporate monopolization of speech. He has suggested that such corporate competition between individual online platforms would result in a “market for rules,” which would allow users to seek networks that have speech and conduct “rule sets” to their liking. 101

Not quite optimists or pessimists, many internet scholars have focused their work on the realities of what the internet is, the harms it does and can create, and the best ways to resolve those harms. Professor Danielle Keats Citron was an early advocate for this approach. She has argued for recognition of cyber civil rights in order to circumvent § 230 immunity without removing the benefits of its protection. 102 Professor Mary Anne Franks has continued this tack, and argues that the nature of online space can amplify speech harms, especially in the context of sexual harassment. 103 Online hate speech, harassment, bullying, and revenge porn have slightly different solutions within these models. Both Citron and Professor Helen Norton have argued that hate speech is now mainstream and should be actively addressed by platforms that have the most power to curtail it. 104 Emily Bazelon argues that the rise of online bullying calls for a more narrow reading of § 230. 105 Citron and Franks respectively suggest either an amendment or a court-created narrowing of § 230 for sites that host revenge porn. 106

This is where we stand today in understanding internet intermediaries: amidst a § 230 dilemma (is it about enabling platforms to edit their sites or about protecting users from collateral censorship?), a First Amendment enigma (what are online platforms for the purposes of speech — a company town, a broadcaster, or an editor?), and conflicting scholarly theories of how best to understand speech on the internet.


103 Mary Anne Franks, Sexual Harassment 2, 71 Md. L. Rev. 655, 678, 681–83 (2012).


105 See generally Emily Bazelon, Sticks and Stones: Defeating the Culture of Bullying and Rediscovering the Power of Character and Empathy (2013); Bazelon, supra note 28.

106 Danielle Keats Citron & Mary Anne Franks, Criminalizing Revenge Porn, 49 Wake Forest L. Rev. 345, 359 n.86 (2014).
Missing from the debate around § 230 is the answer to a simple question: given that these platforms have § 230 immunity, why are they bothering to edit? Administrative law scholarship discusses the forces that motivate private actors to voluntarily self-regulate. Some firms or industries have developed self-regulation alongside government regulation. Others see self-regulation as an optimal form of business and company management. And some decide to self-regulate as an attempt to preempt eventual government regulation. Some of these reasons come to bear on platform motivation, but because of immunity under § 230, most are irrelevant. Instead, through historical interviews and archived materials, Part II argues that platforms have created a voluntary system of self-regulation because they are economically motivated to create a hospitable environment for their users in order to incentivize engagement. This self-regulation involves both reflecting the norms of their users around speech as well as keeping up as much speech as possible. Online platforms also self-regulate for reasons of social and corporate responsibility, which in turn reflect free speech norms. These motivations reflect both the Good Samaritan incentives and collateral censorship concerns underlying § 230.

A question is also missing from the debate about how to classify platforms in terms of First Amendment doctrine: what are major online intermediaries actually doing to regulate content on their sites? The next Part discusses why platforms are making the decisions to moderate along such a fine line, while the following Part demonstrates how platforms moderate content through a detailed set of rules, trained human decisionmaking, and reasoning by analogy, all influenced by a pluralistic system of internal and external actors.

---

109 See generally DENNIS C. KINLAW, CONTINUOUS IMPROVEMENT AND MEASUREMENT FOR TOTAL QUALITY (1992) (describing self-regulation, specifically through the use of total quality management and self-auditing, as the best technique for business management and means of achieving customer satisfaction).
110 See RICHARD L. ABEL, AMERICAN LAWyers 142–57 (1989) (discussing private actors’ decisions to self-regulate in order to avoid potential government regulation).
111 See Citron & Norton, supra note 104, at 1454 (discussing how some intermediaries regulate hate speech because they see it as a threat to profits).
112 Id. at 1455 (discussing how some intermediaries regulate hate speech because they see it as a corporate or social responsibility).
II. Why Govern Well? The Role of Free Speech Norms, Corporate Culture, and Economic Incentives in the Development of Content Moderation

In the earliest days of the internet, the regulations concerning the substance and structure of cyberspace were “built by a noncommercial sector [of] researchers and hackers, focused upon building a network.”\(^\text{113}\) Advances in technology as well as the immunity created for internet intermediaries under § 230 led to a new generation of cyberspace. It included collaborative public platforms like Wikipedia,\(^\text{114}\) but it was also populated largely by private commercial platforms.\(^\text{115}\)

As this online space developed, scholars considered what normative values were being built into the infrastructure of the internet. Lessig ascribed a constitutional architecture to the internet “not to describe a hundred-day plan[, but] instead to identify the values that a space should guarantee. . . . [W]e are simply asking: What values should be protected there? What values should be built into the space to encourage what forms of life?”\(^\text{116}\) Writing five years later in 2004,\(^\text{117}\) Balkin argued that the values of cyberspace are inherently democratic — bolstered by the ideals of free speech, individual liberty, and participation.\(^\text{118}\) Both Lessig and Balkin placed the fate of “free speech values”\(^\text{119}\) and the “freedoms and controls of cyberspace”\(^\text{120}\) in the hands of code and architecture online.\(^\text{121}\) “[A] code of cyberspace, defining the freedoms and controls of cyberspace, will be built,” wrote Lessig.\(^\text{122}\) “About that there can be no debate. But by whom, and with what values? That is the only choice we have left to make.”\(^\text{123}\)

There was not much choice about it, but over the last fifteen years, three American companies — YouTube, Facebook, and Twitter — have

\(^{113}\) LESSIG, supra note 21, at 7.


\(^{115}\) LESSIG, supra note 21, at 7 (describing the second generation of the internet as being “built by commerce”).

\(^{116}\) Id. at 6.

\(^{117}\) As calculated from the first distribution of Lessig’s book, LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

\(^{118}\) See Balkin, supra note 7, at 45–49.

\(^{119}\) Id. at 54.

\(^{120}\) LESSIG, supra note 21, at 6.

\(^{121}\) Specifically, Balkin predicted that free speech values of “participation, access, interactivity, democratic control, and the ability to route around and glom on . . . won’t necessarily be protected and enforced through judicial creation of constitutional rights. Rather, they will be protected and enforced through the design of technological systems — code — and through legislative and administrative schemes of regulation.” Balkin, supra note 7, at 54.

\(^{122}\) LESSIG, supra note 21, at 6.

\(^{123}\) Id.
established themselves as dominant platforms in global content sharing and online speech. These platforms are both the architecture for publishing new speech and the architects of the institutional design that governs it. Because of the wide immunity granted by § 230, these architects are free to choose which values they want to protect — or to protect no values at all. So why have they chosen to integrate values into their platform? And what values have been integrated?

It might first be useful to describe what governance means in the context of these platforms. “The term ‘governance’ is popular but imprecise,” and modern use does not assume “governance as a synonym for government.” Rather, “new governance model[s]” identify several features that accurately describe the interplay between user and platform: a “dynamic” and “iterative” “law-making process”; “norm-generating” “[i]ndividuals”; and “convergence of processes and outcomes.” This is the way in which this Article uses the term “governance.” However, the user-platform relationship departs from even this definition because of its private and centralized but also pluralistically networked nature. And it departs even further from other uses of the term “governance,” including “corporate governance” (describing it as centralized management) and public service definitions of “good governance” (describing states with “independent judicial system[s] and legal framework[s]”).

This Part explores this question through archived material and a series of interviews with the policy executives charged with creating the moderation systems for YouTube and Facebook. It concludes that three

---

124 Each of these platforms can of course be thought of differently. Facebook is primarily categorized as a social network site, see danah m. boyd & Nicole B. Ellison, Social Network Sites: Definition, History, and Scholarship, 13 J. COMPUTER-MEDIATED COMM. 210, 210 (2008); YouTube is seen as video-sharing; and Twitter is seen as both a social network and an RSS newsfeed. But all of these sites have one thing in common: they host, publish, and moderate user-generated content. This Article will look at these platforms in that capacity only.

125 R. A. W. Rhodes, The New Governance: Governing Without Government, 44 POL. STUD. 652, 652 (1996). Indeed, the idea of Facebook as a nation-state or government, in the traditional sense, has been analyzed and dismissed. Anupam Chander, Facebookistan, 90 N.C. L. REV. 1807, 1807 (2012) (concluding “regulatory power [over Facebook] is, de facto, dispersed across a wide array of international actors”). Professor Frank Pasquale has described these platforms as “feudal” or “sovereigns,” FRANK PASQUALE, THE BLACK BOX SOCIETY 140–68, 187–218 (2015) (arguing that terms of service or contracts are inappropriate or ineffective remedies in an essentially “feudal” sphere, id. at 144, and arguing that platforms act as “sovereign[s]” over realms of life, id. at 163, 189), while Professor Rory Van Loo has called them “digital regulators,” Van Loo, supra note 100, at 1267.


127 Id. at 406.

128 Id.

main factors influenced the development of these platforms’ moderation systems: (1) an underlying belief in free speech norms; (2) a sense of corporate responsibility; and (3) the necessity of meeting users’ norms for economic viability.

A. Platforms’ Baseline in Free Speech

Conversations with the people who were in charge of creating the content-moderation regimes at these platforms reveal that they were indeed influenced by the concerns about user free speech and collateral censorship raised in Zeran.

1. Free Speech Norms. — For those closely following the development of online regulation, § 230 and Zeran were obvious foundational moments for internet speech. But at the time, many online commercial platforms did not think of themselves as related to speech at all. As a young First Amendment lawyer in the Bay Area, Nicole Wong was an active witness to the development of private internet companies’ speech policies. In the first few years of widespread internet use, Wong recalled that very few lawyers were focusing on the responsibilities that commercial online companies and platforms might have toward moderating speech. But as most major print newspapers began posting content on websites between 1996 and 1998, the overlap between speech and the internet became more noticeable. Likewise, just as more traditional publishing platforms for speech were finding their place on the internet, new internet companies were discovering that they were not just software companies, but that they were also publishing platforms. At first, Wong’s clients were experiencing speech as only a secondary effect of their primary business, as in the case of Silicon Investor, a day-trading site that was having issues with the content published on its message boards. Others, like Yahoo, were actively recognizing that online speech was an intractable part of their business models. Despite this reality, the transition to thinking of themselves as speech platforms was still slow. “They had just gone public,” Wong said of her representation of early Yahoo. “They had only two lawyers in their legal department. . . . [N]either had any background in First Amendment law or content moderation or privacy. They were corporate

130 Telephone Interview with Nicole Wong, Former Emp., Google (Apr. 2, 2016).
131 Id.
133 Telephone Interview with Nicole Wong, supra note 130.
134 Id.
135 Id.
lawyers.” The problem identified by Wong was that these new internet corporations still thought of themselves as software companies — they did not think about “the lingering effects of speech as part of what they were doing.” In facing these new challenges, Wong had become one of the few people not only in Silicon Valley, but also in the United States, capable of advising on these challenges, with her background in First Amendment doctrine, communications, and electronic privacy.

Wong’s expertise led her to join Google full time in 2004. In October 2006, Google acquired YouTube, the popular online video site, and Wong was put in charge of creating and implementing content-moderation policies. Creating the policies regarding what type of content would be acceptable on YouTube had an important free speech baseline: legal content would not be removed unless it violated site rules. Wong and her content-moderation team actively worked to try to make sure these rules did not result in overcensorship of user speech. One such moment occurred in late December 2006, when two videos of Saddam Hussein’s hanging surfaced on YouTube shortly after his death. One video contained grainy footage of the hanging itself; the other contained video of Hussein’s corpse in the morgue. Both videos violated YouTube’s community guidelines at the time — though for slightly different reasons. “The question was whether to keep either of them up,” said Wong, “and we decided to keep the one of the hanging itself, because we felt from a historical perspective it had real value.” The second video was deemed “gratuitous violence” and removed from the site. A similarly significant exception occurred in June 2009, when a video of a dying Iranian Green Movement protestor shot in the chest and bleeding from the eyes was ultimately kept on YouTube because of its political significance. YouTube’s policies and internal guidelines on violence were altered to allow for the exception. In 2007, a video was uploaded to YouTube of a man being brutally beaten by four men in a cell and was removed for gratuitous violence in violation of

136 Id.
137 Id.
138 For an example of Wong’s insight into these issues, see ELECTRONIC MEDIA AND PRIVACY LAW HANDBOOK (Nicole Wong et al. eds., 2003).
139 Telephone Interview with Nicole Wong, supra note 130; Rosen, supra note 12.
140 Site rules for impermissible content were related to banning content that was otherwise legal but that contained things like graphic violence or overt sexual activity. Buni & Chemaly, supra note 12; see also infra pp. 1632–33.
141 Telephone Interview with Nicole Wong, supra note 130.
142 Id.
143 Buni & Chemaly, supra note 12.
144 Id. It is important to make a distinction between “policies,” which were the public rules posted for users about what content was allowed, and the internal “rules” that sites used to moderate speech. As will be shown in section III.A, infra pp. 1631–35, platforms’ internal rules to moderate content came years before public policies were posted. The internal rules were also more detailed.
YouTube’s community guidelines. Shortly after, however, it was restored by Wong and her team after journalists and protestors contacted YouTube to explain that the video was posted by Egyptian human rights activist Wael Abbas to inform the international community of human rights violations by the police in Egypt.

At Facebook, there was a similar slow move to organize platform policies on user speech. It was not until November 2009, five years after the site was founded, that Facebook created a team of about twelve people to specialize in content moderation. Like YouTube, Facebook hired a lawyer, Jud Hoffman, to head their Online Operations team as Global Policy Manager. Hoffman recalled that, “when I got there, my role didn’t exist.” Hoffman was charged with creating a group separate from operations that would formalize and consolidate an ad hoc draft of rules and ensure that Facebook was transparent with users by publishing a set of “Community Standards.” The team consisted of six people in addition to Hoffman, notably Dave Willner, who had created a first draft of these “all-encompassing” rules, which contained roughly 15,000 words.

At Twitter, the company established an early policy not to police user content, except in certain circumstances, and rigorously defended that right. Adherence to this ethos led to Twitter’s early reputation among social media platforms as “the free speech wing of the free speech

146 Telephone Interview with Nicole Wong, supra note 130.
147 Telephone Interview with Dave Willner, Former Head of Content Policy, Facebook & Charlotte Willner, Former Safety Manager, User Operations, Facebook (Mar. 23, 2016).
149 Id. “Community Standards” is Facebook’s term for its public content-moderation policies. It is important to note that the internal rules created by Dave Willner predated the public Community Standards for the site. The internal rules informed, in part, the creation and substance of Facebook’s public policies.
150 Id.

Our goal is to provide a service that allows you to discover and receive content from sources that interest you as well as to share your content with others. We respect the ownership of the content that users share and each user is responsible for the content he or she provides.

Id. “Because of these principles, we do not actively monitor user’s content and will not censor user content, except in limited circumstances . . . .” Id.
party.”\textsuperscript{152} It also meant that unlike YouTube and Facebook, which actively took on content moderation of their users’ content, Twitter developed no internal content-moderation process for taking down and reviewing content. The devotion to a fundamental free speech standard was reflected not only in what Twitter did not do to police user content, but also in what it did to protect it. Alexander Macgillivray joined Twitter as General Counsel in September 2009, a position he held for four years.\textsuperscript{153} In that time, Macgillivray regularly resisted government requests for user information and user takedown. “We value the reputation we have for defending and respecting the user’s voice,” Macgillivray stated in 2012.\textsuperscript{154} “We think it’s important to our company and the way users think about whether to use Twitter, as compared to other services.”\textsuperscript{155}

A common theme exists in all three of these platforms’ histories: American lawyers trained and acculturated in American free speech norms and First Amendment law oversaw the development of company content-moderation policy. Though they might not have “directly imported First Amendment doctrine,” the normative background in free speech had a direct impact on how they structured their policies.\textsuperscript{156} Wong, Hoffman, and Willner all described being acutely aware of their predisposition to American democratic culture, which put a large emphasis on free speech and American cultural norms. Simultaneously, there were complicated implications in trying to implement those American democratic cultural norms within a global company. “We were really conscious of not just wholesale adopting a kind of U.S. jurisprudence free expression approach,” said Hoffman.\textsuperscript{157} “[We would] try to step back and focus on the mission [of the company].”\textsuperscript{158} Facebook’s mission is to “[g]ive people the power to build community and bring the world closer together.”\textsuperscript{159} But even this, Willner acknowledged, is “not a cultural-neutral mission. . . . The idea that the world

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{152} Josh Halliday, \textit{Twitter’s Tony Wang: “We Are the Free Speech Wing of the Free Speech Party,”} \textsc{The Guardian} (Mar. 22, 2012, 11:57 AM), http://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/QR8B-CW74].
\item \textsuperscript{154} Sengupta, \textit{Twitter’s Free Speech Defender}, supra note 153.
\item \textsuperscript{155} Id.
\item \textsuperscript{156} Telephone Interview with Jud Hoffman, supra note 148.
\item \textsuperscript{157} Id.
\item \textsuperscript{158} Id.
\item \textsuperscript{159} \textit{About, Facebook}, https://www.facebook.com/pg/facebook/about/ [https://perma.cc/3ZV5-MECX].
\end{itemize}
\end{footnotesize}
should be more open and connected is not something that, for example, North Korea agrees with."

2. Government Request and Collateral Censorship Concerns. — Beyond holding general beliefs in the right to users’ free speech, these platforms have also implemented policies to protect user speech from the threat of government request and collateral censorship.

Twitter’s early pushback to government requests related to its users’ content is well documented. In his time as General Counsel, Macgillivray regularly resisted government requests for user information and user takedown. In January 2011, he successfully resisted a federal gag order over a subpoena in a grand jury investigation into Wikileaks. "[T]here’s not yet a culture of companies standing up for users when governments and companies come knocking with subpoenas looking for user data or to unmask an anonymous commenter who says mean things about a company or the local sheriff,” said Wired of Twitter’s resistance to the gag order. “Twitter deserves recognition for its principled upholding of the spirit of the First Amendment." Despite the victory over the gag order, Twitter was eventually forced to turn over data to the Justice Department after exhausting all its appeals.

A similar scenario played out in New York, when a judge ordered Twitter to supply all the Twitter posts of Malcolm Harris, an Occupy Wall Street protestor charged with disorderly conduct. There, too, Twitter lost, but not before full resort to the appeals process.

160 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.
161 This is not to say that collateral censorship issues are not a concern with private platforms’ content-moderation systems. To the contrary, there are also many well-documented instances where platforms have cooperated with government requests for takedown and raised serious collateral censorship concerns. This section simply tries to give an overview of when platforms have proactively sought to avoid these concerns, even though doing so is costly and not necessary under § 230. See Balkin, supra note 11, at 2298–99 (explaining how the government can offer both carrots and sticks to entice private entities to cooperate with it regarding speech regulation); see also, e.g., Emma Llansó, German Proposal Threatens Censorship on Wide Array of Online Services, CTR. FOR DEMOCRACY & TECH.: BLOG (Apr. 7, 2017), https://cdt.org/blog/german-proposal-threatens-censorship-on-wide-array-of-online-services/ [https://perma.cc/W9QT-5MP6] (discussing the dangers of allowing government units to flag issues for takedown using private content-moderation platforms).
163 Id.
164 Id.
165 Sengupta, Twitter’s Free Speech Defender, supra note 153.
166 Id.
Wong also described regularly fighting government requests to take down certain content, collateral censorship, and the problems with applying American free speech norms globally. For example, in November 2006, the Thai government announced that it would block YouTube to anyone using a Thai IP address unless Google removed twenty offensive videos from the site.\textsuperscript{168} While some of the videos “clearly violated the YouTube terms of service,” others simply featured Photoshopped images of the King of Thailand with feet on his head.\textsuperscript{169} In Thailand, insulting the King was illegal and punishable by as much as fifteen years in prison.\textsuperscript{170} Nicole Wong was hard pressed to find the content offensive. “My first instinct was it’s a cartoon. It’s a stupid Photoshop,” she stated, “but then it suddenly became a kind of learning moment for me about international speech standards versus First Amendment speech standards and there was a lot more American First Amendment exceptionalism [in that space] than previously.”\textsuperscript{171} Wong traveled to Thailand to resolve the dispute and was overwhelmed by the popular love she observed in the Thai people for their King. “You can’t even imagine [their love for their King],” she recounted of the trip:

Every Monday literally eighty-five percent of the people show up to work in a gold or yellow shirt and dress\textsuperscript{172} and there’s a historical reason for it: the only source of stability in this country is this King . . . They absolutely revere their King . . . . Someone at the U.S. Embassy described him as a “blend of George Washington, Jesus, and Elvis.” Some people . . . . tears came to their eyes as they talked about the insults to the King and how much it offended them. That’s the part that set me back. Who am I, a U.S. attorney sitting in California to tell them: “No, we’re not taking that down. You’re going to have to live with that.”\textsuperscript{173}

After the trip, Wong and her colleagues agreed to remove the videos within the geographical boundaries of Thailand, with the exception of critiques of the military.\textsuperscript{174}

A few months later, events similar to those in Thailand emerged, but ended in a different result. In March 2007, Turkey blocked access to YouTube for all Turkish users in response to a judge-mandated order.\textsuperscript{175} The judgment came in response to a parody news broadcast that jokingly quipped that the founder of modern Turkey, Mustafa Kemal

\textsuperscript{168} Rosen, supra note 12.
\textsuperscript{169} Id.
\textsuperscript{171} Telephone Interview with Nicole Wong, supra note 130.
\textsuperscript{172} Yellow is the color associated with the King in Thailand. Profile: Thailand’s Reds and Yellows, BBC NEWS (July 13, 2012), http://www.bbc.com/news/world-asia-pacific-13294268 [https://perma.cc/K79R-5AWP] (calling yellow ”the king’s colour”).
\textsuperscript{173} Telephone Interview with Nicole Wong, supra note 130.
\textsuperscript{174} Id.
\textsuperscript{175} Rosen, supra note 12.
Atatürk, was gay.176 As with the King in Thailand, ridicule or insult of Atatürk was illegal in Turkey. Though the video had already been voluntarily removed, Turkey had searched and provided Google with a list of dozens of similarly offensive videos and demanded their takedown.177 Unwilling to meet the blanket demand, Wong and her colleagues at Google found themselves parsing the intricacies of Turkish law on defamation of Atatürk, measuring those standards against the videos highlighted as offensive by the Turkish government, and then offering compromises to ban in Turkey only those videos that they found actually violated Turkish law.178 This seemed to strike an accord for a period of time.179 A little over a year later, however, in June 2007, the Turkish government demanded Google ban access to all such videos not only in Turkey, but worldwide.180 Google refused, and Turkey subsequently blocked YouTube throughout Turkey.181

All three platforms faced the issue of free speech concerns versus censorship directly through platform rules or collateral censorship by government request when a video called *Innocence of Muslims* was uploaded to YouTube.182 Subtitled “The Real Life of Muhammad,” the video depicts Muslims burning the homes of Egyptian Christians, before cutting to “cartoonish” images that paint Muhammad as a bastard, homosexual, womanizer, and violent bully.183 The video’s negative depiction of the Muslim faith sparked a firestorm of outrage in the Islamic world and fostered anti-Western sentiment.184 As violence moved from Libya to Egypt, YouTube issued a statement that while the video would remain posted on the site because the content was “clearly within [its] guidelines,” access to the video would be temporarily restricted in Libya and Egypt.185

At Facebook, the debate between violation of platform guidelines versus concerns over collateral censorship also played out. By the time the video was posted, many of Facebook’s difficulties with hate speech had been distilled into a single rule: attacks on institutions (for example,
countries, religions, or leaders) would be considered permissible content and stay up, but attacks on groups (people of a certain religion, race, or country) would be taken down. In application, this meant that statements like “I hate Islam” were permissible on Facebook, while “I hate Muslims” was not. Hoffman, Willner, and their team watched the video, found no violative statements against Muslims, and decided to keep it on the site. A few weeks later, the Obama Administration called on YouTube to reconsider leaving the video up, in part to quell the violence abroad. Both YouTube and Facebook stuck to their decisions. Reviewing this moment in history, Professor Jeffrey Rosen spoke to the significance of their decisions for collateral censorship: “In this case . . . the mobs fell well outside of U.S. jurisdiction, and the link between the video and potential violence also wasn’t clear . . . Had YouTube made a different decision . . . millions of viewers across the globe [would have been denied] access to a newsworthy story and the chance to form their own opinions.”

The early history and personnel of these companies demonstrate how American free speech norms and concerns over censorship became instilled in the speech policies of these companies. But they also raise a new question: if all three companies had § 230 immunity and all valued their users’ free speech rights, why did they bother curating at all?

B. Why Moderate At All?

These online platforms have broad freedom to shape online expression and a demonstrated interest in free speech values. So why do they bother to create intricate content-moderation systems to remove speech? Why go to the trouble to take down and then reinstate videos of violence like those Wong described? Why not just keep them up in the first place? The answers to these questions lead to the incentives for platforms to minimize online obscenity put in place by the Good Samaritan provision of § 230. Platforms create rules and systems to curate speech out of a sense of corporate social responsibility, but also, more importantly, because their economic viability depends on meeting users’ speech and community norms.

1. Corporate Responsibility and Identity. — Some platforms choose to moderate content that is obscene, violent, or hate speech out of a sense

---

186 Rosen, supra note 176.
187 Id.
189 Id.
190 Rosen, supra note 176.
191 These systems are discussed in detail in Part III, infra pp. 1630–62.
of corporate responsibility. At YouTube, Wong looked to the values of the company in addition to American free speech norms in developing an approach to content moderation. "Not everyone has to be a free-wheeling, free speech platform that is the left wing of the left wing party," she said, referring to Twitter’s unofficial content-moderation policy:

But you get to decide what the tone and tenor of your platform look[] like, and that’s a First Amendment right in and of itself. Yahoo or Google had a strong orientation toward free speech, [and] being more permissive of a wide range of ideas and the way those ideas are expressed, they created community guidelines to set what [users] can come here for, because they want the largest possible audience to join.

Like Wong, Hoffman and Willner considered the mission of Facebook — “to make the world more open and connected” — and found that it often aligned with larger American free speech and democratic values. These philosophies were balanced against competing principles of user safety, harm to users, public relations concerns for Facebook, and the revenue implications of certain content for advertisers. The balance often favored free speech ideals of “leaving content up” while at the same time trying to figure out new approaches or rules that would still satisfy concerned users and encourage them to connect and interact on the platform. "We felt like Facebook was the most important platform for this kind of communication, and we felt like it was our responsibility to figure out an answer to this," said Hoffman.

Likewise, Twitter’s corporate philosophy of freedom of speech justified its failure to moderate content. In recent years, Twitter’s approach has started to change. In a Washington Post editorial, the new General Counsel of Twitter, Vijaya Gadde, used very different rhetoric.

192 See Citron & Norton, supra note 104, at 1455 n.119 (“Such decisions may be justified as a matter of corporate law under the social entity theory of the corporation, which permits corporate decision-makers to consider and serve the interests of all the various constituencies affected by the corporation’s operation.” (citing Lisa M. Fairfax, Doing Well While Doing Good: Reassessing the Scope of Directors’ Fiduciary Obligations in For-Profit Corporations with Non-Shareholder Beneficiaries, 59 WASH. & LEE L. REV. 409, 412 (2002))).

193 Telephone Interview with Nicole Wong, supra note 130.

194 Id.


196 See Telephone Interview with Jud Hoffman, supra note 148; see also Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.

197 Telephone Interview with Jud Hoffman, supra note 148.

198 Id.

199 Id.

200 See supra pp. 1620–21.
than that of her predecessor: “Freedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up,” wrote Gadde. 201 “We need to do a better job combating abuse without chilling or silencing speech.” 202 Over the last two years, the company has slowly made good on its promise, putting a number of policies and tools in place to make it easier for users to filter and hide content they do not want to see. 203

2. Economic Reasons. — Though corporate responsibility is a noble aim, the primary reason companies take down obscene and violent material is the threat that allowing such material poses to potential profits based in advertising revenue. 204 Platforms’ “sense of the bottom-line benefits of addressing hate speech can be shaped by consumers’ — i.e., users’ — expectations.” 205 If a platform creates a site that matches users’ expectations, users will spend more time on the site and advertising revenue will increase. 206 Take down too much content and you lose not only the opportunity for interaction, but also the potential trust of users. Likewise, keeping up all content on a site risks making users uncomfortable and losing page views and revenue. According to Willner and Hoffman, this theory underlies much of the economic rationale behind Facebook’s extensive moderation policies. 207 As Willner stated, “Facebook is profitable only because when you add up a lot of tiny interactions worth nothing, it is suddenly worth billions of dollars.” 208 Wong spoke

202 Id.
203 Kate Klonick, Here’s What It Would Take for Twitter to Get Serious About Its Harassment Problem, VOX (Oct. 25, 2016, 10:50 AM), http://www.vox.com/new-money/2016/10/25/1386648/twitter-harassment-explained [https://perma.cc/VA7M-TRTH]. It is important to note that these methods used by Twitter to maximize free speech by shielding the viewer are really just a type of shadow censorship.
204 See Citron & Norton, supra note 104, at 1454 n.113 (“[T]he traditional ‘shareholder primacy’ view . . . understands the corporation’s primary (and perhaps exclusive) objective as maximizing shareholder wealth.” (first citing Mark J. Roe, The Shareholder Wealth Maximization Norm and Industrial Organization, 149 U. PA. L. REV. 2063, 2065 (2001); then citing A. A. Berle, Jr., For Whom Corporate Managers Are Trustees: A Note, 45 HARV. L. REV. 1365, 1367–69 (1932))).
205 Id.
206 Paul Alan Levy, Stanley Fish Leads the Charge Against Immunity for Internet Hosts — But Ignores the Costs, PUB. CITIZEN: CONSUMER L. & POL’Y BLOG (Jan. 8, 2011), http://pubcit.typepad.com/cdpblog/2011/01/stanley-fish-leads-the-charge-against-immunity-for-internet-hosts-but-ignores-the-costs.html [https://perma.cc/APS9-49BC] (arguing that websites that fail to provide protections against abuse will find “that the ordinary consumers whom they hope to serve will find it too uncomfortable to spend time on their sites, and their sites will lose social utility (and, perhaps more cynically, they know they will lose page views that help their ad revenue)’’); see also Citron & Norton, supra note 104, at 1454 (discussing “digital hate as a potential threat to profits”).
207 Telephone Interview with Jud Hoffman, supra note 148; Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.
208 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.
of the challenge to meet users’ expectations online slightly differently: as platforms attempting to catch up to changing social norms online. 209 Changing expectations about speech are happening both at the platform level, and also at a societal level, said Wong, who referred to the last twenty years of online speech as undergoing a “norm-setting process” that is developing at light speed in comparison to any other kind of publication platform. 210 “What we’re still in the middle of is how do we think about . . . the norms of behavior when what’s appropriate is constantly reiterated,” said Wong. 211 “If you layer over all of that the technology change and the cultural, racial, national, [and] global perspectives, it’s all just changing dramatically fast. It’s enormously difficult to figure out those norms, let alone create policy to reflect them.” 212 Nevertheless, reflecting these rapidly changing norms, and, accordingly, encouraging and facilitating platform interactions — users posting, commenting, liking, and sharing content — is how platforms like Facebook and YouTube have stayed in business and where platforms like Twitter have run into trouble.

Twitter’s transformation from internet hero for its blanket refusal to police users’ content to internet villain happened relatively swiftly. Though public awareness of online hate speech and harassment was already growing, the GamerGate controversy in 2014 raised new levels of global awareness about the issue. 213 As the least policed or rule-based platform, much of the blame fell on Twitter. 214 By 2015, the change in cultural values and expectations began to be reflected in new public standards and policy at Twitter. The site added new language prohibiting “promot[ing] violence against others . . . on the basis of race, ethnicity, national origin, religion, sexual orientation, gender, gender identity, age, or disability” to the Twitter Rules and prohibited revenge

209 Telephone Interview with Nicole Wong, supra note 130.
210 Id.
211 Id.
212 Id.
213 In August of that year, anonymous users targeted a number of women in the gaming industry — including game developers Zoë Quinn, Brianna Wu, and critic Anita Sarkeesian — in a series of harassment campaigns across multiple platforms, including Twitter. Jason Schreier, Thousands Rally Online Against GamerGate, KOTAKU (Oct. 15, 2014, 10:48 AM), https://kotaku.com/thousands-rally-online-against-gamer-gate-1646500492 [https://perma.cc/7E49-9KJA]. The harassment efforts included doxing, as well as rape and death threats. Sarah Kaplan, With #GamerGate, the Video-Game Industry’s Growing Pains Go Viral, WASH. POST (Sept. 12, 2014), http://wapo.st/2EvoOJ7 [https://perma.cc/C4YH-B7J6]. The widespread and graphic nature of the controversy shifted norms and led to many calls on social media platforms to take a more proactive stance against online harassment and hate speech. Schreier, supra.
On December 30, 2015, Twitter published a new set of Twitter Rules — which were largely nothing new, but rather an official incorporation of the separate pages and policies in one place. In January 2016, one Twitter spokesperson described the changes: “Over the last year, we have clarified and tightened our policies to reduce abuse, including prohibiting indirect threats and nonconsensual nude images. Striking the right balance will inevitably create tension, but user safety is critical to our mission at Twitter and our unwavering support for freedom of expression.”

In the mid-1990s, Post presciently wrote about how this interplay between users’ norms around speech and content of online platforms would play out. Post suggested competition between individual online platforms would result in a “market for rules,” which would allow users to seek networks that have “rule sets” to their liking. At least with regard to Twitter, this platform-exit prediction is mostly accurate. Over the last few years, many users unhappy with the policies of Twitter left the platform and favored other platforms like Facebook, Instagram, and Snapchat. As Twitter’s user growth stagnated, many blamed the site’s inability to police harassment, hate speech, and trolling on its site for the slump. In late 2016, Twitter announced a host of new services for users to control their experience online, block hate speech and harassment, and control trolls. Post’s idea of a “market for rules” is an incredibly useful heuristic to understand the history of online content moderation, with two small updates: (1) the history of Twitter reveals a nuance not fully predicted by Post — that is, rather than exit a platform, some users would stay and expect platforms to alter rule sets and policies reactively in response to user pressure; and (2) the “market for rules”

---

215 Jeong, supra note 151 (alterations in original); see also Issie Lapowsky, Why Twitter Is Finally Taking a Stand Against Trolls, WIRED (Apr. 21, 2015, 2:14 PM), https://www.wired.com/2015/04/twitter-abuse/[https://perma.cc/R43VK].


217 Jeong, supra note 151.

218 Post, supra note 101, at para. 42.

219 See Jeff Dunn, Here’s How Slowly Twitter Has Grown Compared to Facebook, Instagram, and Snapchat, BUS. INSIDER (Feb. 10, 2017, 6:14 PM), http://www.businessinsider.com/twitter-vs-facebook-snapchat-user-growth-chart-2017-2 [https://perma.cc/PF6U-GGPC] (assuming arguendo that growth in market alone cannot account for the slow growth of users on Twitter as compared to the growth of users on social media platforms like Snapchat, Instagram, and Facebook).


221 Klonick, supra note 205.
paradigm mistakes the commodity at stake in online platforms. The commodity is not just the user, but rather it is the content created and engaged with by a user culture. In this sense there is no competition between social media platforms themselves, as Post suggests, because they are complementary, not substitute, goods.

Whether rooted in corporate social responsibility or profits, the development of platforms’ content-moderation systems to reflect the normative expectations of users is precisely what the creation of the Good Samaritan provision in § 230 sought. Moreover, the careful monitoring of these systems to ensure user speech is protected can be traced to the free speech concerns of § 230 outlined in Zeran. The answer to the dilemma of what § 230 protects — immunity for good actors creating decency online or protection against collateral censorship — seems not to be an either/or answer. Rather, both purposes seem to have an essential role to play in the balance of private moderation of online speech.

With this new knowledge about the motivations behind platforms’ content-moderation systems, we can then ask the next question in the debate over internet intermediaries: how are platforms actually moderating? The answer to this question, explored in the next Part, is essential to understanding how platforms should — or should not — be understood for the purposes of First Amendment law.

III. HOW ARE PLATFORMS GOVERNING? THE RULES, PROCESS, AND REVISION OF CONTENT-MODERATION SYSTEMS

Much of the analysis over how to categorize online platforms with respect to the First Amendment is missing a hard look at what these platforms are actually doing and how they are doing it. In part, this is because the private content-moderation systems of major platforms like Facebook, Twitter, and YouTube are historically opaque. This Part seeks to demonstrate how these systems actually work to moderate online speech. In doing this, Part III looks at the history of how content-moderation systems changed from those of standards to those of rules, how platforms enforce these rules, and how these rules are subject to change. Many of these features bear remarkable resemblance to heuristics and structures familiar in legal decisionmaking. Despite these similarities, platform features are best thought of not in terms of First Amendment law.

222 See Balkin, supra note 7, at 4–6.

223 Moreover, Post’s free-market idea of user exit is also challenged by current studies. In an ongoing project, the Electronic Frontier Foundation has worked to document and present evidence of the negative psychological impact that leaving — either by choice or by banning — certain social media platforms can have on users. See Submit Report, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/submit-report [https://perma.cc/25NK-LGA2] (offering a platform for users to report erroneous or unjust account deactivations). These studies support the theory Lessig describes in Code: Version 2.0, in which he proffers that leaving an internet platform is more difficult and costly than expected. See LESSIG, supra note 21, at 288–90.
Amendment doctrine — as reflecting the role of a state actor, a broadcaster, or a newspaper editor — but in terms of a private self-regulatory system to govern online speech.

A. Development of Moderation: From Standards to Rules

When Dave Willner joined a small team to specialize in content moderation in November 2009, no public “Community Standards” existed at Facebook. Instead, all content moderation was based on one page of internal “rules” applied globally to all users. Willner recalled that the moderation policies and guidance for enforcing them were limited.\(^{224}\) “The [policy] guidance was about a page; a list of things you should delete: so it was things like Hitler and naked people. None of those things were wrong, but there was no explicit framework for why those things were on the list.”\(^{225}\) Willner’s now-wife Charlotte was also working at Facebook doing customer service and content moderation and had been there for a year before Dave joined.\(^{226}\) She described the ethos of the pre-2008 moderation guidelines as “if it makes you feel bad in your gut, then go ahead and take it down.”\(^{227}\) She recalled that the “Feel bad? Take it down” rule was the bulk of her moderation training prior to the formation of Dave’s group in late 2008.\(^{228}\) Wong described a similar ethos in the early days at YouTube, especially around efforts to know when to remove graphic violence from the site. Speaking of reinstating the 2007 video of the Egyptian protestors being brutally beaten,\(^{229}\) Wong said: “It had no title on it. It wasn’t posted by him. . . . I had no way of knowing what it was and I had taken something down that had real significance as a human rights document. So we put it back up. And then we had to create another exception to the no-violence rule.”\(^{230}\) Though both Wong and the Willners used the term “rule” in describing these prescriptions for takedown, a more precise term for these early guidelines might be “standard.” In legal theory, the “rules-standards conflict” describes the battle between two formal resolutions for legal controversy.\(^{231}\) An example of a standard is “don’t drive too fast.” An

\(^{224}\) Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.

\(^{225}\) Id.

\(^{226}\) Id.

\(^{227}\) Id.

\(^{228}\) Id.

\(^{229}\) See supra pp. 1619–20.

\(^{229}\) Telephone Interview with Nicole Wong, supra note 130.

example of a rule is a speed limit set at sixty-five miles per hour. There are trade-offs to picking one as the formal solution over the other. Standards are often “restatements of purpose” or values, but because they are often vague and open ended, they can be “subject to arbitrary and/or prejudiced enforcement” by decisionmakers. This purposive approach, however, can also mean that standards are enforced precisely and efficiently and can be more accommodating to changing circumstances. Rules, on the other hand, have the issues reverse to those of standards. Rules are comparatively cheap and easy to enforce, but they can be over- and underinclusive and, thus, can lead to unfair results. Rules permit little discretion and in this sense limit the whims of decisionmakers, but they also can contain gaps and conflicts, creating complexity and litigation.

Whichever approach is used, a central point is that the principles formalized in rules and standards are rooted in the social norms and values of a community. Standards are more direct analogues of values or purpose but “require[] that the enforcing community . . . come to some consensus on the meaning of a value term.” Rules are more distant from the norms they are based on and “do not depend on ongoing dialogue to gain dimension or content . . . even by someone who shares no sense of community with his fellows.”

The development at YouTube and Facebook from standards to rules for content moderation reflects these trade-offs. A simple standard against something like gratuitous violence is able to reach a more tailored and precise measure of justice that reflects the norms of the community, but it is vague, capricious, fact dependent, and costly to enforce.

This can be seen at YouTube, which in mid-2006 employed just sixty workers to review all video that had been flagged by users for all reasons. For violations of terms of service, one team of ten, deemed the Safety, Quality, and User Advocacy Department, or SQUAD, worked in shifts “around the clock” to keep YouTube from “becoming a shock site.” That team was given a one-page bullet-point list of standards that instructed on removal of things like animal abuse, videos showing blood, visible nudity, and pornography. A few months later, in the

232 KELMAN, supra note 231, at 40 (emphasis omitted).
233 Id. at 41.
234 Id. at 40.
235 See id. at 40–47.
237 KELMAN, supra note 231, at 61.
238 Id. at 62.
239 Buni & Chemaly, supra note 12.
240 Id.
241 Id.
fall of 2006, the YouTube list turned into a six-page booklet drafted with input from the SQUAD, Wong, and other YouTube lawyers and policy executives.  

Five years later, in 2011, the volume of uploaded video to YouTube had more than doubled in size, making delicate, precise decisions less feasible. In addition, the content-moderation team had expanded and been outsourced. Accordingly, the more individually tailored standards against gratuitous violence had slowly been replaced by precise rules, which were easier and less costly to enforce. Moderators were given a booklet with internal rules for content moderation. This booklet was regularly annotated and republished with changes to moderation policies and rules. Many of these new rules were drafted as “exceptions” to rules. Eventually, a more detailed iterative list of rules and their exceptions largely replaced the standards-based approach of earlier years.

Similar to the experience at YouTube, Facebook eventually abandoned the standards-based approach as the volume of user-generated content increased, the user base diversified, and the content moderators globalized. Dave Willner was at the helm of this transition. Though Facebook had been open globally for years, Willner described much of the user base during his early days there as still relatively homogenous — “mostly American college students” — but that was rapidly changing as mobile technology improved and international access to the site grew. Continuing to do content moderation from a single list of banned content seemed untenable and unwieldy. Instead, Willner set about changing the entire approach:

In the early drafts we had a lot of policies that were like: “Take down all the bad things. Take down things that are mean, or racist, or bullying.” Those are all important concepts, but they’re value judgments. You have to be more granular and less abstract than that. Because if you say to forty college students [content moderators], “delete all racist speech,” they are not going to agree with each other about what’s racist.

Eliminating standards that evoked nonobservable values, feelings, or other subjective reactions was central to Willner’s new rulebook for moderation. Instead, he focused on the implicit logic of the existing page of internal guidelines and his experience and extrapolated from them to

---

242 Id.

243 On May 1, 2009, YouTube had twenty hours of video upload per minute; by May 1, 2011, forty-eight hours of video were uploaded per minute. See Mark R. Robertson, 500 Hours of Video Uploaded to YouTube Every Minute [Forecast], TUBULAR INSIGHTS (Nov. 13, 2015), http://tubularinsights.com/hours-minute-uploaded-youtube/ [https://perma.cc/A9Q7-N3VM].

244 Buni & Chemaly, supra note 12.

245 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147; see also Buni & Chemaly, supra note 12.

246 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.
create objective rules. The first draft of these “all-encompassing” rules was written largely by Willner in 2009 and contained roughly 15,000 words. The end goal was consistency and uniformity: to get the same judgment on a piece of content, regardless of who was moderating it.

Exactly “who” was moderating the content changed significantly in January 2009, when Facebook opened its office in Dublin and first started outsourcing its content moderation through consulting groups. Before then, most moderators worked in Palo Alto and were similar to Facebook’s main user base — “homogenous college students.” The shift to outsourced moderation continued when a new community operations team was set up in Hyderabad, India. Around the same time, Hoffman joined Facebook’s team as Global Policy Manager with the goal of formalizing and consolidating the rules Willner had started to draft, and ensuring that Facebook was transparent with users by publishing a set of public rules in the form of “Community Standards.”

Hoffman and Willner worked together to transform the early ad hoc abuse standards into operational internal rules for content moderators, a document that today is over eighty pages long. This movement from standards to rules was “ultimately a form of technical writing,” said Willner. “You cannot tell people to delete photos with ugly clothes in them. You have to say ‘delete photos with orange hats in them.’” For Willner, some of the hardest parts of defining categories, elements, and distinctions came in moderating art and nudity. For Hoffman, it was more difficult to create rules around hate speech. “We couldn’t make a policy that said ‘no use of the N-word at all,’” he recalled, describing the difficulty in policing racial slurs. “That could be completely insensitive to the African American community in the United States. But you also don’t want it used as hate speech. So it’s

\[\text{247 Id.}\]
\[\text{248 Id.}\]
\[\text{249 Id.}\]
\[\text{250 Id.}\]
\[\text{251 Id.}\]
\[\text{252 Id. “Community Standards” is Facebook’s term for its public content-moderation policies. It is important to note that the internal rules created by Willner predated the public Community Standards for the site. In fact, it was the internal rules that informed, in part, the creation and substance of Facebook’s public policies.}\]
\[\text{253 Id.}\]
\[\text{254 Id.}\]
\[\text{255 Id.}\]
\[\text{256 “Art doesn’t exist as a property of an image. There are no art pixels that you can find in images we think are classy or beautiful or uplifting. . . . But what we realized about art was that [moderation] questions about art weren’t about art itself, it was about art being an exception to an existing restriction. . . . [S]o the vast majority of art is fine. It’s when you’re talking about things that might meet the definition of nudity or racism or violence that people think are important.” Id.}\]
\[\text{257 Telephone Interview with Jud Hoffman, supra note 148.}\]
almost impossible to turn that into an objective decision because context matters so much.\textsuperscript{258} The answer was to turn context into a set of objective rules. In evaluating whether speech was likely to provoke violence, for example, Hoffman and his team developed a four-part test to assess credible threats: time, place, method, and target.\textsuperscript{259} If a post specified any three of these factors, the content would be removed, and if appropriate, authorities notified.\textsuperscript{260}

Content moderation at YouTube and Facebook developed from an early system of standards to an intricate system of rules due to \((1)\) the rapid increase in both users and volume of content; \((2)\) the globalization and diversity of the online community; and \((3)\) the increased reliance on teams of human moderators with diverse backgrounds. The next section discusses enforcement of these rules.

\textbf{B. How the Rules Are Enforced: Trained Human Decisionmaking}

Content moderation happens at many levels. It can happen before content is actually published on the site, as with ex ante moderation, or after content is published, as with ex post moderation. These methods can be either \textit{reactive}, in which moderators passively assess content and update software only after others bring the content to their attention, or \textit{proactive}, in which teams of moderators actively seek out published content for removal. Additionally, these decisions can be \textit{automatically} made by software or \textit{manually} made by humans.\textsuperscript{261} The majority of

\begin{footnotesize}
\begin{itemize}
  \item \textsuperscript{258} \textit{Id.}
  \item \textsuperscript{259} Univ. of Hous. Law Ctr., \textit{UH Law Center and the ADL Present Racists, Bigots and the Law on the Internet}, YOUTUBE (Oct. 10, 2012), \url{https://youtu.be/aqqvYPyr6cT?list=UU3rht1s6oKV8PnWrd47KQ} [https://perma.cc/Q7SF-TYNM][recording of Jud Hoffman, Glob. Policy Manager, Facebook].
  \item \textsuperscript{260} \textit{Id.} Many situations, however, were lacking in context. Online bullying was the type of issue that often arose with insufficient background. As Hoffman described:
    
    There is a traditional definition of bullying — a difference in social power between two people, a history of contact — there are elements. But when you get a report of bullying, you just don’t know. You have no access to those things. So you have to decide whether you’re going to assume the existence of some of those things or assume away the existence of some of those things. Ultimately what we generally decided on was, “if you tell us that this is about you and you don’t like it, and you’re a private individual not a public figure, we’ll take it down.” Because we can’t know whether all these other things happened, and we still have to make those calls. But I’m positive that people were using that function to game the system. . . . I just don’t know if we made the right call or the wrong call or at what time.
    
    Telephone Interview with Jud Hoffman, supra note 148. Hoffman’s description also demonstrates two major drawbacks to using rules rather than standards. A blanket rule against bullying can simultaneously result in people manipulating a rule to “walk the line” and also result in permissible content being mistakenly removed. \textit{Id.}
  \item \textsuperscript{261} \textit{See} James Grimmelmann, \textit{The Virtues of Moderation}, 17 YALE J.L. \\& TECH. 42, 63–70 (2015) (describing how moderation systems operate differently along several lines — automatic or manual, transparent or secret, ex ante or ex post, and centralized or decentralized). Professor James Grimmelmann’s taxonomy, while foundational, speaks more generally to all of internet moderation.
\end{itemize}
\end{footnotesize}
this section focuses on ex post reactive content moderation, specifically looking at the implementation of rules with respect to human decisionmaking, pattern recognition, and professionalization of judgment.

1. Ex Ante Content Moderation. — When a user uploads a video to Facebook, a message appears: “Processing Videos: The video in your post is being processed. We’ll send you a notification when it’s done and your post is ready to view.” Ex ante content moderation is the process that happens in this moment between “upload” and publication. The vast majority of this moderation is an automatic process run largely through algorithmic screening without the active use of human decisionmaking.

An example of content that can be moderated by these methods is child pornography, which can reliably be identified upon upload through a picture-recognition algorithm called PhotoDNA. Under federal law, production, distribution, reception, and possession of an image of child pornography is illegal, and as such, sites are obligated to remove it. A known universe of child pornography — around 720,000 illegal images — exists online. By converting each of these images to grayscale, overlaying a grid, and assigning a numerical value to each

rather than content-publishing platforms specifically. In the context of speech, the distinction between ex ante and ex post is especially important, in that it determines whether moderation is happening before or after publication. Of secondary concern is whether content is being moderated through reactive or proactive measures. Finally, the ultimate means of reaching decisions, whether through software or humans, is descriptively helpful, but less legally significant.


263 Because ex ante content moderation happens before publication takes place, it is the type of prior restraint that scholars like Balkin are concerned with. See generally Balkin, supra note 11. Of the two automatic means of reviewing and censoring content — algorithm and geoblocking — geoblocking is of more concern for the purposes of collateral censorship and prior restraint. In contrast, algorithms are currently used to remove illegal content like child pornography or copyright violations. But see Rebecca Tushnet, Power Without Responsibility: Intermediaries and the First Amendment, 76 G.E.O. WASH. L. REV. 986, 1003–05 (2008) (noting that the Digital Millennium Copyright Act’s notice-and-takedown provisions give platforms no incentive to investigate and therefore “suppress critical speech as well as copyright infringement,” id. at 1003).


265 See 18 U.S.C. §§ 2251–2252A (2012). It is important to remember that § 230 expressly states that no internet entity has immunity from federal criminal law, intellectual property law, or communications privacy law. 47 U.S.C. § 230(e)(2) (2012). This means that every internet service provider, search engine, social networking platform, and website is subject to thousands of laws, including child pornography laws, obscenity laws, stalking laws, and copyright laws. Id.

square, researchers were able to create a “hash,” or signature, that remained even if the images were altered. As a result, platforms can determine whether an image contains child pornography in the microseconds between upload and publication. Geoblocking is another form of automatic ex ante moderation. Unlike PhotoDNA, which prevents the publication of illegal content, geoblocking prevents both publication and viewing of certain content based on a user’s location. As happened in the controversy over the Innocence of Muslims video, geoblocking usually comes at the request of a government notifying a platform that a certain type of posted content violates its local laws.

Of course, algorithms do not decide for themselves which kind of content they should block from being posted. Content screened automatically is typically content that can reliably be identified by software and is illegal or otherwise prohibited on the platform. This universe of content that is automatically moderated ex ante is regularly evaluated and updated through iterative software updates and machine learning. For example, in a similar fashion to PhotoDNA, potential copyright violations can be moderated proactively through software like Content ID. Developed by YouTube, Content ID allows creators to give their content a “digital fingerprint” so it can be compared against other uploaded content. Copyright holders can also flag already-published copyright violations through notice and takedown. These two systems work together, with user-flagged copyrighted material eventually added to ContentID databases for future proactive review. This mix of proactive, manual moderation and informed, automatic ex ante moderation is also evident in the control of spam. All three platforms (and most internet companies, generally) struggle to control spam postings on their sites. Today, spam is mostly blocked automatically from publication through software. Facebook, Twitter, and YouTube, however, all feature mechanisms for users to report spam manually. Ex ante screening software is iteratively updated to reflect these flagged spam sources.

---

267 Id.
268 Id., supra note 264.
269 See supra pp. 1624–25; see also, e.g., Telephone Interview with Nicole Wong, supra note 130.
272 How Content ID Works, supra note 270.
2. Ex Post Proactive Manual Content Moderation. — Recently, a form of content moderation that harkens to the earlier era of AOL chat rooms has reemerged: platforms proactively seeking out and removing published content. Currently, this method is largely confined to the moderation of extremist and terrorist speech. As of February 2016, dedicated teams at Facebook have proactively removed all posts or profiles with links to terrorist activity. Such efforts were doubled in the wake of terrorist attacks. This is an important new development affecting content moderation, which seeks to strike an ever-evolving balance between competing interests: ensuring national security and maintaining individual liberty and freedom of expression. While a topic worthy of deep discussion, it is not the focus of this paper.

3. Ex Post Reactive Manual Content Moderation. — With the exception of proactive moderation for terrorism described above, almost all user-generated content that is published is reviewed reactively, that is, through ex post flagging by other users and review by human content moderators against internal guidelines. Flagging — alternatively called reporting — is the mechanism provided by platforms to allow users to express concerns about potentially offensive content. The adoption by social media platforms of a flagging system serves two main functions: (1) it is a “practical” means of reviewing huge volumes of content, and (2) its reliance on users serves to legitimize the system when platforms are questioned for censoring or banning content.

Facebook users flag over one million pieces of content worldwide every day. Content can be flagged for a variety of reasons, and the vast majority of items flagged do not violate the Community Standards of Facebook. Instead content flags often reflect internal group conflicts or disagreements of opinion. To resolve the issue, Facebook created a new reporting “flow” — the industry term to describe the sequence of screens users experience as they make selections — that encourages users to resolve issues themselves rather than report them for review to


275 Id.


277 Id. at 412.

278 See Buni & Chemaly, supra note 12; Telephone Interview with Monika Bickert, Head of Glob. Policy Mgmt., Facebook & Peter Stern, Head of Policy Risk Team, Facebook (Jan. 19, 2016).

279 The Trust Engineers, RADIOLAB (Feb. 9, 2015, 8:01 PM), https://www.radiolab.org/story/trust-engineers/ [https://perma.cc/9C4N-SJRW].
Users reporting content first click a button to “Report/Mark as Spam,” which then quickly guides users to describe their report in terms like “Hate Speech,” “Violence or Harmful Behavior,” or “I Don’t Like This Post.” Some types of reports, such as harassment or self-harm, guide users to the option of “social reporting” — a tool that “enables people to report problematic content not only to Facebook, but also directly to their friends to help resolve conflicts.” To enhance the response time of content moderation, the reporting flow also has the instrumental purpose of triaging flagged content for review. This makes it possible for Facebook to immediately prioritize certain content for review and, when necessary, notify authorities of emergency situations like suicide, imminent threats of violence, terrorism, or self-harm. Other content, like possible hate speech, nudity, pornography, or harassment, can be queued into less urgent databases for general review.

After content has been flagged to a platform for review, the precise mechanics of the decisionmaking process become murky. The “army” of content moderators and “[t]he details of moderation practices are routinely hidden from public view,” write Catherine Buni and Soraya Chemaly. Social media companies do not publish details of their internal content moderation guidelines; no major platform has made such guidelines public. These internal guidelines also change much more frequently than the public Terms of Service or Community Standards. Focusing largely on Facebook, except where specified, the next section seeks to illuminate this process by integrating previously published information together with interviews of content moderators and platform internal guidelines. The system of people making the decisions will be examined first followed by a review of the internal guidelines that inform that decisionmaking process.

(a) Who Enforces the Rules? — When content is flagged or reported, it is sent to a server where it awaits review by a human content moderator. At Facebook, there are three basic tiers of content moderators: “Tier 3” moderators, who do the majority of the day-to-day reviewing

---

281 Id.
283 Id.
284 Univ. of Hous. Law Ctr., supra note 259 (Jud Hoffman speaking).
285 Id.
286 Buni & Chemaly, supra note 12.
287 Id.
288 Skype Interviews with Kumar S. (Jan. 29–Mar. 9, 2016); Skype Interviews with Selahattin T. (Mar. 2–Mar. 11, 2016); Skype Interview with Jagruti (Jan. 26, 2016) [hereinafter Content Moderator Interviews]. These content moderators were Tier 3 workers based in India and Eastern Europe and provided background on what the process looks like from the perspective of a content moderator.
of content; “Tier 2” moderators, who supervise Tier 3 moderators and review prioritized or escalated content; and “Tier 1” moderators, who are typically lawyers or policymakers based at company headquarters.289

In the early days, recent college graduates based in the San Francisco Bay Area did much of the Tier 3 content moderation.290 Today, most platforms, including Facebook, either directly employ content-moderation teams or outsource much of their content-moderation work to companies like oDesk (now Upwork), Sutherland, and Deloitte.291 In 2009, Facebook opened an office in Dublin, Ireland, that had twenty dedicated support and user-operations staff.292 In 2010, working with an outsourcing partner, Facebook opened a new office in Hyderabad, India, for user support.293

Today, Tier 3 moderators typically work in “call centers”294 in the Philippines, Ireland, Mexico, Turkey, India, or Eastern Europe.295 Within Facebook, these workers are called “community support” or “user support teams.”296 When working, moderators will log on to computers and access the server where flagged content is awaiting review.297 Tier 3 moderators typically review material that has been flagged as a lower priority by the reporting flow. At Facebook, for example, this includes, in part, reports of nudity or pornography; insults or attacks based on religion, ethnicity, or sexual orientation, inappropriate or annoying content, content that is humiliating, or content that advocates violence to a person or animal.298

Tier 2 moderators are typically supervisors of Tier 3 moderators or specialized moderators with experience judging content. They work both remotely (many live in the United States and supervise groups that

289 Telephone Interview with J.L., Tier 2 Moderator, Facebook (Mar. 11, 2016). J.L. was a Tier 2 moderator based in the Eastern United States.

290 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147; Telephone Interview with Sasha Rosse, Manager of Glob. Outsourcing, Facebook (May 16, 2016); Buni & Chemaly, supra note 12.


292 Telephone Interview with Sasha Rosse, supra note 290.

293 Id.

294 Buni & Chemaly, supra note 12.

295 Content Moderator Interviews, supra note 288; Telephone Interview with Sasha Rosse, supra note 290; Chen, supra note 291; Chen, supra note 12.

296 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147; Telephone Interview with Jud Hoffman, supra note 148.

297 Content Moderator Interviews, supra note 288; Telephone Interview with Sasha Rosse, supra note 290.

298 Telephone Interview with J.L., supra note 289.
are internationally based) and locally at call centers. Tier 2 moderators review content that has been prioritized, like imminent threats of violence, self-harm, terrorism, or suicide. This content comes to Tier 2 directly through the reporting flow or by being identified and escalated to Tier 2 by Tier 3 moderators. Tier 2 moderators also review certain randomized samples of Tier 3 moderation decisions. In order to ensure the accuracy of moderation, Facebook and other platforms have a certain amount of built-in redundancy: the same piece of content is often given to multiple Tier 3 workers. If the judgment on the content varies, the content is reassessed by a Tier 2 moderator.

Tier 1 moderation is predominantly performed at the legal or policy headquarters of a platform. At Facebook, for example, a Tier 3 worker could be based in Hyderabad, a Tier 2 supervisor could be based in Hyderabad, or remotely in a place like Dublin, but a Tier 1 contact would be based in Austin, Texas, or the San Francisco Bay Area. "There were not many levels between the boots-on-ground moderator and Menlo Park," stated one former Tier 2 supervisor who had worked at Facebook until 2012, speaking on the condition of anonymity. "If I had doubts on something, I’d just send it up the chain."

Recently, issues of scaling this model have led platforms to try new approaches to who enforces the rules. At YouTube, a new initiative was launched in late 2016 called the Heroes program, which deputizes users to actively participate in the content-moderation process in exchange for perks such as “access to exclusive workshops and sneak preview product launches.” Similarly, after a video of the murder of an elderly man in Cleveland stayed up for over an hour on Facebook, Zuckerberg announced the company would hire 3000 additional content moderators, increasing the size of the content-moderation team by two-thirds.

(b) How Are the Rules Enforced? — As previously discussed, the external policy — or Community Standards — provided to the public is not the same as the internal rulebook used by moderators when trying to assess whether content violates a platform’s terms of service. An analysis of the internal guidelines reveals a structure that in many ways

---

299 Id.; Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.
300 Telephone Interview with J.L., supra note 289.
301 Id.
302 Id.
replicates the decisionmaking process present in modern jurisprudence. Content moderators act in a capacity very similar to that of a judge: moderators are trained to exercise professional judgment concerning the application of a platform’s internal rules and, in applying these rules, moderators are expected to use legal concepts like relevance, reason through example and analogy, and apply multifactor tests.

(i) Training. — Willner and Hoffman’s development of objective internal rules at Facebook was a project that became an essential element in the shift to content-moderation outsourcing made in early 2010. While Facebook’s Community Standards were applied globally, without differentiation along cultural or national boundaries, content moderators, in contrast, came with their own cultural inclinations and biases. In order to ensure that the Community Standards were enforced uniformly, it was necessary to minimize content moderators’ application of their own cultural values and norms when reviewing content and instead impose Facebook’s. The key to all of this was providing intensive in-person training on applying the internal rules. “It all comes down to training,” stated Sasha Rosse, who worked with Willner to train the first team in Hyderabad:

I liked to say that our goal was [to have a training system and rules set] so I could go into the deepest of the Amazon, but if I had developed parameters that were clear enough I could teach someone that had no exposure to anything outside of their village how to do this job.

305 Facebook outsourced only a small subset of reports in 2010. Most of the content-moderation work was still being performed by full-time employees and contractors in Hyderabad, Austin, and Palo Alto. Email from Jud Hoffman, Former Glob. Policy Manager, Facebook (Aug. 18, 2016) (on file with author).

306 It is worth noting why Facebook has this policy. According to Willner, in writing the internal rules and the Community Standards, Facebook:

realized that the nature of the product made regional rules untenable. There are no “places” in Facebook — there are just people with different nationalities, all interacting in many shared forums. Regional rules would make cross-border interactions and communities largely incoherent and moderation very hard if not impossible. For example, if a Greek user insults Atatürk and a Turkish user reports it, whose rules apply? Telephone Interview with Dave Willner & Charlotte Willner, supra note 147.

307 Though often referred to as “neutral,” Facebook’s values and norms — and the rules that attempted to reflect them — were distinctly American. See supra section II.A, pp. 1618–25.

308 Telephone Interview with Sasha Rosse, supra note 290. Despite the internal rules and training, cultural biases still crept into moderation, especially when judging subjective content. For example, in 2010 and 2011, the Facebook content-policy team was still struggling to refine its guidelines as it simultaneously began to train moderators in India. Rules on nudity were relatively cut because nudity could in large part be reduced to observable characteristics that were either present or not in content. But harder questions arose regarding the Facebook rules banning certain kinds of sexualized content: a person could be entirely clothed, but in a highly sexual position. At some point in the training process in India, a group of workers were given a list of observable rules about a picture that made it impermissibly sexual, but at the bottom of the rules there was a more general “Feel Bad” standard: if you feel like something is otherwise sexual or pornographic, take it
Training moderators to overcome cultural biases or emotional reactions in the application of rules to facts can be analogized to training lawyers or judges. In the law, training lawyers and judges through law school and practice bestows a “specialized form of cognitive perception — what Karl Llewellyn called ‘situation sense’ — that reliably focuses their attention on the features of a case pertinent to its valid resolution.”

Professor Dan Kahan calls this “professional judgment,” but it might also be called “pattern recognition” after Professor Howard Margolis’s study of expert and lay assessments of risk, or even likened to the rapid, instinctual categorization used by chicken sexers, expert workers whose entire job is to determine the sex of baby chickens a day or two after the chickens hatch. Regardless of the label, training content moderators involves a repetitive process to “override” cultural or emotional reactions and replace them with rational “valid” resolutions.

Recent studies show that professionalized judgment can thwart cognitive biases, in addition to increasing attention to relevant information and reliable application of rules. In a series of experiments, Kahan asked judges, lawyers, and law students with various political inclinations to assess legal problems that were “designed to trigger unconscious political bias in members of the general public.” Despite the presence of irrelevant but polarizing facts, judges, and to a lesser degree, lawyers, were largely in agreement in deciding legal cases presented to them in the study. In contrast, law students and members of the general public reliably made decisions in keeping with their personal political views when presented with politically polarizing information. Replication of the study expanded these findings beyond mere political ideologies to more general “cultural cognition,” that is, the “unconscious influence of
individuals’ group commitments on their perceptions of legally consequential facts.317

The experiments by Kahan and his co-authors demonstrate empirically what Facebook learned through experience: people can be trained in domain-specific areas to overcome their cultural biases and to apply rules neutrally. Just as this truth is an essential part of the legal system, it is an essential part of Facebook’s moderation system.

(ii) Similarities to American Law and Legal Reasoning. — Before applying law to facts, a judge must first determine which facts are relevant. Procedural rules like the Federal Rules of Evidence acknowledge that the inclusion of certain information may unfairly exploit decisionmakers’ biases and emotions and, thus, provide guidance on how to exclude information from review.318 At Facebook, the internal rules used by content moderators, or “Abuse Standards,” similarly contain extensive guidance on what “relevant” content a moderator should review in assessing a report.319

Once a moderator has followed the procedural rules to narrow the relevant content to be reviewed, the actual Abuse Standards — or rules — must be applied. These start with a list of per se bans on content.320 In Abuse Standards 6.2, these per se bans on content are lists of rules split into nine somewhat overlapping categories.321 But as is typical of a rules-based approach, these lists contain as many exceptions as they do rules. In “Graphic Content,” listed violations include any “[p]oaching of animals” as well as “[p]hotos and digital images showing internal organs, bone, muscle, tendons, etc.,” while “[c]rushed heads,

317 Kahan et al., supra note 313, at 851.
318 See Kahan et al., supra note 309, at 365.
320 AS 6.1, supra note 319; AS 6.2, supra note 319.
321 Those categories are “Sex and Nudity,” “Illegal Drug Use,” “Theft Vandalism and Fraud,” “Hate Content,” “Graphic Content,” “IP Blocks and International Compliance,” “Self Harm,” “Bullying and Harassment,” and “Credible Threats.” AS 6.2, supra note 319, at 4. Among the twelve items under “Sex and Nudity,” for example, are “[a]ny OBVIOUS sexual activity, even if naked parts are hidden from view by hands, clothes or other objects. Cartoons/art included. Foreplay allowed (kissing, groping, etc.) even for same-sex individuals” and “[p]eople ‘using the bathroom.’” Id.
limbs, etc. are ok as long as no insides are showing.” Likewise, “mere depiction” of some types of content — “hate symbols” like swastikas, or depictions of Hitler or Bin Laden — are automatic violations, “unless the caption (or other relevant content) suggests that the user is not promoting, encouraging or glorifying the [symbol].”

Some more complicated types of speech borrow from American jurisprudence for the structure of their rules. Under “Hate Content,” a chart provides examples of “Protected Categories” and counsels moderators to mark “content that degrades individuals based on the . . . protected categories” as a violation. A second chart on the page demonstrates how the identification of the type of person — ordinary persons, public figures, law enforcement officers, and heads of state — as well as their membership in a protected group will factor into the permissibility of the content. All credible threats are to be escalated regardless of the “type of person.” These examples demonstrate the influence of American jurisprudence on the development of these rules. Reference to “Protected Categories” is similar to the protected classes of the Civil Rights Act of 1964. The distinction between public and private figures is reminiscent of First Amendment, defamation, and invasion of privacy law. The emphasis on credibility of threats harkens to the balance between free speech and criminal law.

Beyond borrowing from the law substantively, the Abuse Standards borrow from the way the law is applied, providing examples and analogies to help moderators apply the rules. Analogical legal reasoning, the method whereby judges reach decisions by reasoning through analogy —

322 Id.
323 Id. at 8.
324 Id. at 5 (including race, ethnicity, national origin, religion, sex, gender identity, sexual orientation, disability, and serious disease as protected categories).
325 Id. An empty threat against a public figure like Paul McCartney is permissible, but an empty threat against a head of state like President Barack Obama should be removed. Any type of content about a law enforcement officer — empty threat, credible threat, negative reference, cyberbullying, and attacks with hate symbols — is a violation under the Abuse Standards, as is any kind of attack based on being a victim of sexual assault. Id.
326 “For safety and legal reasons, we consider threats credible if they:
1. Target heads of state or specific law enforcement officers . . .
2. Contain 3/4 details: time, place, method, specific target (not impossible to carry out)
3. Target people with a history of assassination attempts
4. Include non-governmental bounties (promising earthly and heavenly rewards for a target’s death)

Id. at 7.
329 See, e.g., Brett A. Sokolow et al., The Intersection of Free Speech and Harassment Rules, 38 HUM. RTS. 19, 19 (2011).
between cases, is a foundation of legal theory. Though the use of example and analogy plays a central role throughout the Abuse Standards, the combination of legal rule and example in content moderation seems to contain elements of both rule-based legal reasoning and analogical legal reasoning. For example, after stating the rules for assessing credibility, the Abuse Standards give a series of examples of instances that establish credible or noncredible threats. “I’m going to stab (method) Lisa H. (target) at the frat party (place),” states Abuse Standards 6.2, demonstrating a type of credible threat that should be escalated. “I’m going to blow up the planet on new year’s eve this year” is given as an example of a noncredible threat. Thus, content moderators are not expected to reason directly from prior content decisions as in common law — but the public policies, internal rules, examples, and analogies they are given in their rulebook are informed by past assessments.

In many ways, platforms’ evolution from “gut check” standards to more specific rules tracks the evolution of the Supreme Court’s doctrine defining obscenity. In Jacobellis v. Ohio, Justice Stewart wrote that he could not “intelligibly” define what qualified something as obscene, but famously remarked, “I know it when I see it.” Both Charlotte Willner, at Facebook, and Nicole Wong, at Google, described a similar intuitive ethos for removing material in the early days of the platforms’ content-moderation policies. Eventually, Facebook’s and YouTube’s moderation standards moved from these standards to rules. Likewise, over a series of decisions, the Court attempted to make the criteria for obscenity more specific — in Miller v. California, the Court issued a


331 This is especially true in the case of introducing new rules or policies to moderators. For example, Abuse Standards 6.2 introduces a “fresh policy” on sexually explicit language and sexual solicitation, and lists thirteen examples of content that should be removed or kept up under the policy. AS 6.2, supra note 319, at 6. In Abuse Standards 6.1, an entire page is devoted to samples of pictures that fall in or out of the various bans on sex and nudity, cartoon bestiality, graphic violence, animal abuse, or Photoshopped images. AS 6.1, supra note 319, at 4.

332 Id. at 7.

333 Id.


335 Id. at 197 (Stewart, J., concurring).

336 Telephone Interview with Dave Willner & Charlotte Willner, supra note 147; Telephone Interview with Nicole Wong, supra note 130.

three-part test to evaluate whether state statutes designed to regulate obscene materials were sufficiently limited.\footnote{Id. at 24.} None of the tests created by the Court, however, comes close to the specificity of the facts and exceptions used by platforms today.

To summarize, knowledge about the training of content moderators and Abuse Standards 6.1 and 6.2 tells us much about how the rules are enforced in content-moderation decisions. Content moderators act in a capacity very similar to that of judges: (1) like judges, moderators are trained to exercise professional judgment concerning the application of a platform’s internal rules; and (2) in applying these rules, moderators are expected to use legal concepts like relevancy, reason through example and analogy, and apply multifactor tests.

4. Decisions, Escalations, and Appeals. — At Facebook, Tier 3 moderators have three decisionmaking options regarding content: they can “confirm” that the content violates the Community Standards and remove it, “unconfirm” that the content violates Community Standards and leave it up, or escalate review of the content to a Tier 2 moderator or supervisor.\footnote{AS 6.1, \textit{supra} note 319; AS 6.2, \textit{supra} note 319.} The Abuse Standards describe certain types of content requiring mandatory escalations, such as: child nudity or pornography, bestiality, credible threats, self-harm, poaching of endangered animals, Holocaust denial, all attacks on Atatürk, maps of Kurdistan, and burning of Turkish Flags.\footnote{AS 6.2, \textit{supra} note 319, at 4.} If a moderator has decided to ban content, a Facebook user’s content is taken down, and she is automatically signed off of Facebook. When the user next attempts to sign in, she will be given the following message: “We removed the post because it doesn’t follow the Facebook Community Standards.”\footnote{Screenshot of Facebook Removal Notice, M.E.ME (June 13, 2017, 5:38 AM), https://me.me/i/29-21-01-am-facebook-we-removed-something-you-posted-we-removed-15347765 [https://perma.cc/2BHA-936B].} When she clicks “Continue,” the user is told: “Please Review the Community Standards: We created the Facebook Community Standards to help make Facebook a safe place for people to connect with the world around them. Please read the Facebook Community Standards to learn what kinds of posts are allowed on Facebook.”\footnote{Screenshot of Facebook Community Standards Notice, M.E.ME (May 1, 2017, 1:09 PM), https://me.me/i/?-30-am-pao-94-facebook-please-review-the-community-standards-13463100 [https://perma.cc/LNQ9-HHRV].} The user then clicks “Okay” and is allowed to log back in. At Facebook, users who repeatedly have content removed face a gradual intensification of punishment: two removed posts in a certain amount of time, for example, might mean your account is suspended for twenty-four hours. Further violations of community standards can result in total bans. At YouTube, moderators had

\footnote{339 Id. at 24.  
a slightly different set of options for each piece of content: “Approve” let a video remain; “Racy” gave the video an 18+ year-old rating; “Reject” allowed a video to be removed without penalizing the poster; and finally, “Strike” would remove the video and issue a penalty to the poster’s account.344

The ability of an individual user to appeal a decision on content takedown, account suspension, or account deletion varies widely between the three major platforms. Facebook allows an appeal of the removal of only a profile or page — not individual posts or content.345 To initiate an appeal process, a user’s account must have been suspended.346 Appeals are reviewed by the Community Operations teams on a rolling basis and sent to special reviewers.347 In contrast, at YouTube, account suspensions, “strikes” on an account, and content removal are all appealable.348 Video strikes can be appealed only once, and if a decision to strike is upheld, there is a sixty-day moratorium on the appeal of any additional strikes.349 An appeal also lies if an account is terminated for repeated violations.350 At Twitter, any form of action related to the Twitter Rules can be appealed.351 Users follow instructions on the app itself or provided in an email sent to notify users that content has been taken down.352 Twitter also includes an intermediary level of removal called a “media filter” on content that might be sensitive.353 Rather than totally remove the content, the platform requires users to click through a warning in order to see the content.354 Appeals are handled by support teams that, when possible, will use specialized team members to review culturally specific content.355

C. System Revision and the Pluralistic System of Influence

Facebook’s Abuse Standards do more than shed light on substantive rules on speech or the mechanisms behind its decisionmaking. They also demonstrate that the internal rules of content moderation are iteratively revised on an ongoing basis, and much more frequently than the external public-facing policy. This can be seen on the first page of Abuse

344 Buni & Chemaly, supra note 12.
345 How to Appeal, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/resources/how-to-appeal [https://perma.cc/VM9B-QzK9].
346 Id.
347 Telephone Interview with J.L., supra note 289.
348 How to Appeal, supra note 345.
349 Id.
350 Id.
351 Id.
352 Id.
353 Id.
354 Id.
355 Id.
Standards 6.1, titled “Major changes since Abuse Standards 6.0,” which contains a bulleted list of amendments and alterations to the previous set of rules.356 The list is divided into groupings of roughly related policy changes.357 “Added sexual language and solicitation policy,” states the first bullet.358 Halfway down the page after “Sex & Nudity issues clarified” is a bulleted section beginning “Graphic violence policies updated as follows.”359 In Abuse Standards 6.2, there are fewer updates, summarized broadly under one bullet point, “Policy Changes”:

- Graphic Content with respect to animal insides
- Threshold and considerations for credible threats
- Caricatures of protected categories
- Depicting bodily fluids
- Screenshots or other content revealing personal information
- PKK versus Kurdistan flags
- Updated policy on photo-shopped images

The differences between these two versions demonstrate that internal policies and the rules that reflect them are constantly being updated. This is because Facebook is attempting, in large part, to rapidly reflect the norms and expectations of its users.

But how are platforms made aware of these “dramatically fast” changing global norms such that they are able to alter the rules? This section discusses four major ways platforms’ content-moderation policies are subject to outside influence: (1) government request, (2) media coverage, (3) third-party civil society groups, and (4) individual users’ use of the moderation process.

This multi-input content-moderation system is a type of pluralistic system.360 Under the ideal theory, a pluralistic system consists of many

---

357 Id.
358 Id.
359 Id. “No exceptions for news or awareness-related context for graphic image depictions — confirm all such content; [human/animal abuse subject to clear involvement/enjoyment/approval/encouragement by the poster [should be confirmed]; [even fake/digital images of graphic content should be confirmed, but hand-drawn/cartoon/art images are ok.” Id. (emphasis omitted).
361 Telephone Interview with Nicole Wong, supra note 130.
diverse external factions of equal strength competing to influence a neutral government. In a perfect world, the competition between these minority factional interests serves to maintain equilibrium and representation of ideas in a democratic society. But in practice, pluralism can be far from ideal. This section discusses these interests and their potential democratic conflicts in the context of outside influence on platforms’ content-moderation policies.

1. Government Requests. — Lessig describes the architecture of the internet — or constitution — as built by an “invisible hand, pushed by government and by commerce.” Lessig does not describe these two forces as separate, but rather tandem in their effect. Thus far, this Article has principally focused on the commercial side of this dynamism, but platform architecture has also been informed by and subject to government interference. This interference can be through the more direct need to comply with local laws and jurisdictions, or by the more subtle influences of government lobbying and requests.

The previous examples of the Thai King, Atatürk, and Innocence of Muslims illustrate how platforms have either conformed their policies, modified their policies, or rejected policy changes following government request. At YouTube, material would be removed within a country only if it violated the laws of that country — whether or not it was a violation was determined by YouTube’s own lawyers. If content was found to be in violation of a country’s laws, a new policy would be issued and geoblocks put in place to prevent access to that content.

363 HELD, supra note 362, at 57–64.
364 Freeman, supra note 15, at 560 (“Although conscious of capture, the theory envisions this pathology as limited to agencies, and as correctable, presumably by democratizing the agency decision-making process to include numerous interest groups. In this sense, interest representation reveals a lingering optimism about the democratic potential of pluralism, when properly structured.”). But see Richard B. Stewart, The Reformation of American Administrative Law, 88 HARV. L. REV. 1667, 1713 (1975) (discussing how some interests like those of a regulated party may be overrepresented in agency government).
365 See Freeman, supra note 15, at 560 (discussing the threat capture poses to democratic ideals of pluralism); see also Margot E. Kaminski, When the Default Is No Penalty: Negotiating Privacy at the NTIA, 93 DENV. L. REV. 925 (2016) (examining instances in which openness to participation by interest groups did not result in meaningful participation); David Thaw, Enlightened Regulatory Capture, 89 WASH. L. REV. 329 (2014) (examining instances in which regulatory capture by a concentrated interest group can be beneficial).
366 For an excellent and more thorough discussion of the potential effects of government requests on online platforms and free speech, see Llansó, supra note 161.
367 LESSIG, supra note 21, at 4.
368 See section II.A.1–2, supra pp. 1618–25 (detailing how Wong established geoblocking within Thailand for some types of content — determined by YouTube — that ridiculed the King; how Wong established geoblocking within Turkey for some types of content — determined by YouTube — that disparaged Atatürk; and how Facebook and YouTube refused requests of the government to remove Innocence of Muslims, and instead kept it up as permissive under their own moderation rules and standards).
369 Telephone Interview with Nicole Wong, supra note 130.
within that country. Similar agreements were reached regarding depictions of Atatürk in Turkey. At Facebook, however, content is not geoblocked but removed globally if international compliance requires. Other types of content with specific geographic sensitivities, like Holocaust denial focusing on hate speech, attacks on Atatürk, maps of Kurdistan, and burning of Turkish flags, are required to be escalated.

Twitter maintains a policy that it will take down posts only on request and only if they violate a country’s laws. This policy has occasionally been at odds with Twitter’s more unofficial tactic of vigorously and litigiously protecting free speech. Compromises have been reached, however, without sacrificing one for the other: in 2012, when India demanded Twitter remove a number of accounts that were fueling religious dissent, the company removed roughly half of the problematic accounts, but did so on the grounds that they violated Twitter’s own policies for impersonation. In other contexts, such as the requests for takedown in Egypt and Turkey, particularly during periods of revolution, Twitter has refused to capitulate to any government requests, and governments have consequently blocked the platform.

Recently, however, platforms have been criticized for increasingly acquiescing to government requests, especially in the distribution of user information to police. Platforms have also begun cooperating more proactively in response to the increased use of social media by the Islamic State of Iraq and Syria (ISIS) to recruit members and encourage terrorism. Over the last few years, all three sites have agreed to general requests from the United States and the United Nations to remove content related to ISIS or terrorism. As discussed briefly in section III.B,
Facebook now maintains a team that is focused on terrorism-related content and helps promote “counter speech” against such groups. The team actively polices terrorist pages and friend networks on the site. No posts from known terrorists are allowed on the site, even if the posts have nothing to do with terrorism. “If it’s the leader of Boko Haram and he wants to post pictures of his two-year-old and some kittens, that would not be allowed,” said Monika Bickert, Facebook’s head of global policy management. As Facebook has become more adept at and committed to removing such terrorism-related content, that content has moved to less restrictive platforms like Twitter. In just a four-month period in 2014, ISIS supporters used an estimated 46,000 Twitter accounts, though not all were active simultaneously. Just before the dissemination of pictures of American journalist James Foley’s beheading, the platform in 2015 began taking a different approach. In early 2016, Twitter reported that it had suspended 125,000 accounts related to ISIS.

2. Media Coverage. — The media do not have a major role in changing platform policy per se, but when media coverage is coupled with either (1) the collective action of users or (2) a public figure’s involvement, platforms have historically been responsive.

An early high-profile example of media catalyzing collective action occurred around a clash between Facebook’s nudity policy and breastfeeding photos posted by users. As early as 2008, Facebook received criticism for removing posts that depicted a woman breastfeeding. The specifics of what triggered removal changed over time. The changes came, in part, after a campaign in the media and in pages on Facebook itself staged partly by women who had their content removed. Similar policy changes occurred after public outcry over


381 Andrews & Seetharaman, supra note 274.
382 Telephone Interview with Monika Bickert & Peter Stern, supra note 279.
385 Andrews & Seetharaman, supra note 274.
388 Id.
Facebook’s “real name” policy, removal of a gay kiss, censoring of an 1886 painting that depicted a nude woman, posting of a beheading video, and takedown of photos depicting doll nipples.

The vulnerability of platforms to public collective action via the media is an important statement on platforms’ democratic legitimacy. The media can serve to lend “civility” to individual speech, and render it more capable of effecting change. Though, of course, Facebook, Twitter, and YouTube are not democratic institutions, they arise out of a democratic culture. Thus, users’ sense that these platforms respond to collective publicized complaints can impact their trust and use of the company. In a recent survey of Americans who received some of their news from social media, eighty-seven percent used Facebook and trusted the platform more than YouTube and Twitter. These numbers held even following reports that Facebook used politically biased algorithms to post news in its “Trending Topics.” While it is impossible to attribute Facebook’s high user base and trust entirely to its responsiveness to

---


393 Asher Moses, “Facebook Relents on Doll Nipples Ban,” SYDNEY MORNING HERALD (July 12, 2010), [https://perma.cc/PY8-SD7A].


395 See Miami Herald Publ’g Co. v. Tornillo, 418 U.S. 241, 249 (1974) (describing how the press is “enormously powerful and influential in its capacity to manipulate popular opinion and change the course of events”), Robert Post, “Participatory Democracy as a Theory of Free Speech: A Reply,” 97 VA. L. REV. 617, 624 (2011) (“Public opinion could not create democratic legitimacy if it were merely the voice of the loudest or the most violent. . . . Public opinion can therefore serve the cause of democratic legitimacy only if it is at least partially formed in compliance with the civility rules that constitute reason and debate.”).

396 How People Decide What News to Trust on Digital Platforms and Social Media, AM. PRESS INST. (Apr. 17, 2016, 10:30 AM), [https://perma.cc/SUZ8-4X3D].

397 Russell Brandom, “After Trending Topics Scandal, Users Still Mostly Trust Facebook,” THE VERGE (May 18, 2016, 6:00 AM), [https://perma.cc/768D-4PP6].
media outcry, the platform’s unique history of altering its policies in response to such complaints has likely fostered its user base. Though ideally democratic, the media can work within this pluralist system to disproportionately favor people with power over the individual users. A series of recent events demonstrate this concept. In September 2016, a well-known Norwegian author, Tom Egeland, posted a famous and historical picture on his Facebook page. The photo of a nine-year-old Vietnamese girl running naked following a napalm attack (“Napalm Girl”) was a graphic but important piece of photo journalism from the Vietnam War. It also violated the terms of service for Facebook. The photo was removed, and Egeland’s account was suspended. In reporting on the takedown, Espen Egil Hansen, the editor-in-chief and CEO of Aftenposten, a Norwegian newspaper, also had the picture removed. Norwegian Prime Minister Erna Solberg also posted the image and had it removed. In response, Hansen published a “letter” to Zuckerberg on Aftenposten’s front page. The letter called for Facebook to create a better system to prevent censorship. Hours later, COO Sheryl Sandberg stated that the company had made a mistake and promised the rules would be rewritten to allow the photo. The responsiveness of Facebook would have been more admirable if this had been the first instance of the Napalm Girl photo ever being censored on the site. But instead, it was likely only one of thousands of times the photo had been removed. To the best of my knowledge, however, all prior instances had failed to happen to a famous author, political world

398 By this I mean power in every sense: power from money, political clout, media access, access to people that work at platforms, celebrity status, a substantial number of followers or friends, or as a verified user.


400 The photo was likely removed because of the nudity, not because it was child pornography. See Kjetil Malkenes Hovland & Deepa Seetharaman, Facebook Backs Down on Censoring “Napalm Girl” Photo, WALL ST. J. (Sept. 9, 2016, 3:07 PM), http://on.wsj.com/2bYZiNR [https://perma.cc/SP8M-UQ5D].

401 Id.


403 Hovland & Seetharaman, supra note 400.

404 See Hansen, supra note 402.


406 Online Chat with Dave Willner, Former Head of Content Policy, Facebook (Sept. 10, 2016).
leader, or the editor-in-chief of a newspaper — and thus, the content had never been reinstated.

Sometimes the speech of powerful people is not just restored upon removal; it is kept up despite breaking the platform policies. In late October, a source at Facebook revealed that Zuckerberg held a Town Hall meeting with employees to discuss why many of then-candidate Donald Trump’s more controversial statements had not been removed from the site even though they violated the hate speech policies of the company.407 “In the weeks ahead, we’re going to begin allowing more items that people find newsworthy, significant, or important to the public interest — even if they might otherwise violate our standards,” senior members of Facebook’s policy team wrote in a public post.408 Despite that, many employees continued to protest that Facebook was unequally and unfairly applying its terms of service and content-moderation rules.

3. Third-Party Influences. — For a number of years, platforms have worked with outside groups to discuss how best to construct content-moderation policies. One of the first such meetings occurred in 2012, when Stanford Law School invited many of these platforms to be part of a discussion about online hate speech.409 In April of that year, roughly two dozen attendees — including ask.fm, Facebook, Google, Microsoft, Quizlet, Soundcloud, Twitter, Whisper, Yahoo, and YouTube410 — met to discuss the “challenge of enforcing . . . community guidelines for free speech” between platforms that have “very different ideas about what’s best for the Web.”411 The best practices that came out of these meetings were issued at the conclusion of months of meetings of the Working Group on Cyberhate and were published on the Anti-Defamation League’s (ADL) website in a new page called “Best Practices for Responding to Cyberhate” in September 2014.412 The page “urge[d] members of the Internet Community, including providers, civil society, the legal community and academia, to express their support for this effort and to publicize their own independent efforts to counter cyberhate.”413

Civil society and third-party groups had and continue to have an impact on the policies and practices of major social media platforms.

408 Id.
409 Rosen, supra note 176.
410 Despite the anonymity, the make-up of the group can be estimated from those industry members that signed the best practices at the culmination of the workshops. See Best Practices for Responding to Cyberhate, ANTI-DEFAMATION LEAGUE, http://www.adl.org/combating-hate/cyber-safety/best-practices/ [https://perma.cc/KHS4-PZKE].
411 Rosen, supra note 176.
413 Best Practices for Responding to Cyberhate, supra note 410.
Sit-downs and conversations sponsored by groups like ADL have pushed the creation of industry best practices. Influence also occurs on a smaller scale. “We have a relationship with them where if we flag something for them, they tend to know that it’s serious, that they should look sooner rather than later,” stated a member of one third-party anti-hate speech group speaking anonymously.414 But such a relationship isn’t exclusive to just organized advocates or established groups. Reporter and feminist Soraya Chemaly recounts directly emailing Sandberg in 2012 regarding graphic Facebook pages about rape and battery of women. “She responded immediately,” says Chemaly, “and put us in touch with the head of global policy.”415 Facebook actively encourages this type of engagement with civil society groups, government officials, and reporters. “If there’s something that the media or a government minister or another group sees that they’ve reported and we haven’t taken it down, we want to hear about it,” said Bickert.416 “We’ve been very proactive in engaging with civil society groups all over the world so that we can get a better understanding of the issues affecting them.”417

In terms of impacting policy, the Working Group on Cyberhate, which was formed in 2012 by the Inter-Parliamentary Coalition for Combating Anti-Semitism and the group of industry leaders and stakeholders at Stanford,418 continues to exert influence on the platforms. The group regularly meets to try to tailor platform guidelines to strike the correct balance between freedom of expression and user safety.419 Other groups, like the Electronic Frontier Foundation (EFF), have a slightly less amicable working relationship with these platforms and exist as more like watchdogs than policy collaborators. Launched in 2012, EFF’s site, onlinecensorship.org, works to document when user content is blocked or deleted by providing an online tool where users can report such incidents.420 “Onlinecensorship.org seeks to encourage companies to operate with greater transparency and accountability toward their users as they make decisions that regulate speech,” states the site's

415 Telephone Interview with Soraya Chemaly, Director, Women’s Media Ctr. Speech Project (May 28, 2016); see also Christopher Zara, Facebook Rape Campaign Ignores Critical Twitter: Boycott Threats from #FBrape Get Advertisers’ Attention, INT’L BUS. TIMES (May 24, 2013, 4:26 PM), http://www.ibtimes.com/facebook-rape-campaign-ignores-twitter-boycott-threats-fbrape-get-advertisers-1278999 [https://perma.cc/A5VT-TKCA].
416 Telephone Interview with Monika Bickert & Peter Stern, supra note 279.
417 Id.
418 Best Practices for Responding to Cyberhate, supra note 410.
420 Who We Are, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/about/who-we-are [https://perma.cc/FzL2-YQH6].
About Page.421 “By collecting these reports, we’re . . . looking . . . to build an understanding of how the removal of content affects users’ lives. Often . . . the people that are censored are also those that are least likely to be heard. Our aim is to amplify those voices and help them to advocate for change.”422 The recent, largely opaque, cooperation between content platforms and government to moderate speech related to terrorism is also an issue of concern for EFF, which has urged such groups “not to ‘become agents of the government.’”423 EFF’s director of International Freedom of Expression, Jillian York, said, “I think we have to ask if that’s the appropriate response in a democracy.”424 “While it’s true that companies legally can restrict speech as they see fit, it doesn’t mean that it’s good for society to have the companies that host most of our everyday speech taking on that kind of power.”

4. Change Through Process. — Beyond outside influences, much of the change in moderation policy and guidelines comes simply from the process of moderation. As new situations arise during moderation, platforms will both tweak current policy as well as develop new rules. “People will do everything on the internet,” said Jud Hoffman.426 “Every day you will encounter something new. . . . The difficulty was making sure we were [reacting] fast enough to address the immediate situations that were causing us to consider [changing our approach], but also being thoughtful enough that we weren’t flip-flopping on that particular issue every week.”427 Once the team had come to a conclusion about the “trade-offs” for a new policy, the additions would be disseminated in the new guidelines, which would then be distributed as updates to moderators.428 Many of these judgments continue to be difficult to make, such as, for example, Nicole Wong’s story of removal from YouTube of the beating of an Egyptian dissident. The video was restored once its political significance was understood. “You might see an image that at first blush appears disturbing, yet in many cases it is precisely that sort of power image that can raise consciousness and move people to take action and, therefore, we want to consider very, very seriously the possibility of leaving it up,” said Peter Stern, head of the Policy Risk Team at Facebook.429 “We want people to feel safe on Facebook, but that doesn’t always mean they’re going to feel comfortable, because they may

422 Id.
423 Andrews & Seetharaman, supra note 274.
424 Greenberg, supra note 383.
425 Id.
426 Telephone Interview with Jud Hoffman, supra note 148.
427 Id.
428 Id.
429 Telephone Interview with Monika Bickert & Peter Stern, supra note 279.
be exposed to images that are provocative or even disturbing. We want to leave room for that role to be played as well."

In recent years, Facebook’s approach to altering its policy has been less passive than simply waiting for new types of content to filter through the system. “We’re trying to look beyond individual incidents where we get criticism, to take a broader view of the fabric of our policies, and make sure that we have mitigated risks arising from our policies as much as we can,” said Stern. “This means looking at trends . . . at what people within the company are saying . . . be it reviewers or people who are dealing with government officials in other countries. We regularly take this information and process it and consider alterations in the policy.”

D. Within Categories of the First Amendment

In light of this new information about how platforms work, how would the First Amendment categorize online content platforms: are they state actors under *Marsh*, broadcasters under *Red Lion* and *Turner*, or more like newspaper editors under *Tornillo*?

Of these, only finding platforms to be state actors would confer a First Amendment obligation — a result that is both unlikely and normatively undesirable. In finding state action, the Court in *Marsh* was particularly concerned with who regulated the municipal powers, public services, and infrastructure of the company town — the streets, sewers, police, and postal service. Subsequent courts have concluded that these facts bear on whether “the private entity has exercised powers that are ‘traditionally the exclusive prerogative of the State.’” This Article has detailed how platforms have developed a similar infrastructure to regulate users’ speech through detailed rules, active and passive moderation, trained human decisionmaking, reasoning by analogy, and input from internal and external sources. Yet this similarity, while perhaps moving in a direction which might someday evoke *Marsh*, is not yet enough to turn online platforms into state actors under the state action

---

430 Id.
431 Id.
432 Id.
434 Blum v. Yaretsky, 457 U.S. 991, 1005 (1982) (quoting Jackson v. Metro. Edison Co., 419 U.S. 345, 353 (1974)). This test is known as the exclusive public function test. If the private entity does not exercise such powers, a court must consider whether “the private party has acted with the help of or in concert with state officials.” *McKeesport Hosp. v. Accreditation Council for Graduate Med. Educ.*, 24 F.3d 519, 524 (3d Cir. 1994). The final factor is whether “[t]he State has so far insinuated itself into a position of interdependence with [the acting party] that it must be recognized as a joint participant in the challenged activity.” *Krynicky v. Univ. of Pittsburgh*, 742 F.2d 94, 98 (3d Cir. 1984) (quoting *Burton v. Wilmington Parking Auth.*, 365 U.S. 715, 725 (1961)).
doctrine. In part, this is because while platforms have an incredible
governing system to moderate content and perform a vast number of
other services which might someday be considered “municipal,” they are
far from “exclusive” in their control of these rights. As the presence
of three major sites for posting this content demonstrates, Facebook,
YouTube, and Twitter do not have sole control over speech generally,
only speech on their sites.

The Court’s recent ruling in Packingham, however, could signal a
shift that might change this calculus. If the Court is concerned with
questions of access in order to exercise constitutionally protected rights,
these sites’ ability to remove speakers — and the lack of procedure or
transparency in doing so — might be of central importance. Still, finding
platforms to be state actors seems a long way off and would require
a very expansive interpretation of Marsh’s current doctrine. Even
should the facts necessary to achieve this interpretation come to pass,
the normative implications of such a result make it unlikely. Interpre-
ting online platforms as state actors, and thereby obligating them to pre-
serve the First Amendment rights of their users, would not only explicitly conflict with the purposes of § 230, but would also likely create an
internet nobody wants. Platforms would no longer be able to remove
obscene or violent content. All but the very basest speech would be
explicitly allowed and protected — making current problems of online
hate speech, bullying, and terrorism, with which many activists and
scholars are concerned, unimaginably worse. This alone might be all
that is needed to keep platforms from being categorized as state actors.

If these platforms are not state actors, the question of defining them
under the First Amendment becomes more complicated. Considering
online content providers to be editors like those in Tornillo, for instance,
would grant them special First Amendment protection. While platforms’ omnipresent role seems to be moving them beyond the world of “editors,” Packingham’s new labeling of platforms as “forums” makes dismissing this categorization slightly more difficult. In Tornillo, the Court held that a newspaper was “more than a passive receptacle or conduit for news, comment, and advertising. The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues and public officials — whether fair or unfair — constitute the exercise of editorial control and judgment.”439 Thus, in order not to “dampen[] the vigor and limit[] the variety of public debate,”440 the Court found the newspaper in Tornillo to have rights equivalent to a speaker under the First Amendment.441 At first blush, this analogy seems appealing. As seen above, like the Miami Herald, Facebook, YouTube, and Twitter are not “passive . . . conduit[s] for news, comment, and advertising.” These platforms have intricate systems for controlling the content on their sites. For the content that stays up — like a newspaper determining what space to allot certain issues — platforms also have intricate algorithms to determine what material a user wants to see and what material should be minimized within a newsfeed, homepage, or stream. But a central piece is missing in the comparison to an editorial desk: platforms do not actively solicit specific types of content, unlike how an editorial desk might solicit reporting or journalistic coverage. Instead, users use the site to post or share content independently. Additionally, platforms play no significant role — yet442 — in determining whether content is true or false or whether coverage is fair or unfair. As Willner summarized: “This works like a Toyota factory, not a newsroom.”443 Accordingly, while platforms might increasingly be compared to editors as their presence continues to expand in online discourse, they are still far from constituting editors under Tornillo.444

Perhaps the increasingly apt analogy is — even though the Court in Reno explicitly excluded it — to compare platforms to broadcasters, and then perhaps even to public utilities or common carriers.445 In Reno,

440 Id. at 257 (citing N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279 (1964)).
441 Id. at 258.
443 Klonick, supra note 399.
444 It is worth noting that Wong and others frequently referred to platforms as possessing their own First Amendment rights to create the type of platform they wanted. This argument stems from Tornillo, but it is more ambitious than any rights currently reflected in the doctrine.
the Court explicitly differentiated the internet from broadcast media because the former lacks scarcity, invasiveness, and a history of government regulation.\footnote{Reno v. ACLU, 521 U.S. 844, 868 (1997).} Excepting the lack of historical regulation around the internet, much has changed online since 1998 in terms of internet scarcity and invasiveness. In the years since \textit{Reno}, the hold of certain platforms has arguably created scarcity — if not of speech generally, undoubtedly of certain mediums of speech that these platforms provide. Certainly too, the internet is now more invasive in everyday life than television is — in fact, today, the internet actively threatens to supplant television and broadcasting,\footnote{See, e.g., \textit{Cutting the Cord}, \textit{The Economist} (July 16, 2016), https://www.economist.com/news/business/21702177-television-last-having-its-digital-revolution-moment-cutting-cord [https://perma.cc/N6HC-QE7K].} and the rise in smartphones and portable electronic technology makes the internet and its platforms ubiquitous. Perhaps most convincingly, in the underlying \textit{Red Lion} decision, the Court argued that “[w]ithout government control, the medium would be of little use because of the cacaphony [sic] of competing voices, none of which could be clearly and predictably heard.”\footnote{Red Lion Broad. Co. v. FCC, 395 U.S. 367, 376 (1969).} The recent scourge of online fake news, scamming, and spam makes this seemingly anachronistic concern newly relevant.

As for public utilities or common carriers regulation, the argument has long been applied at the most basic level of the internet to answer concerns over possible politicization of internet service providers\footnote{GOLDSMITH & WU, supra note 100, at 72–74.} that act as content-neutral conduits for speech. But this argument fails for platforms, because they are inherently \textit{not} neutral — indeed the very definition of “content moderation” belies the idea of content neutrality. Nevertheless, the “essential” nature of these private services to a public right — and the prominence of a few platforms which hold an increasingly powerful market share — evinces concerns similar to those of the people who are arguing for regulation of telephone or broadband services.

A few other analogies that implicate the First Amendment might also apply, but they all fail to match the scope and scale of the speech happening on online platforms. Platforms’ use of rule sets to govern speech is reminiscent of “speech codes” used by universities to constrain the speech rights of the student body. But private universities are not truly full-fledged forums — not in the way that California and New Jersey treat shopping malls,\footnote{See, e.g., PruneYard Shopping Ctr. v. Robins, 447 U.S. 74, 78 (1980).} and not in the way that platforms have become forums for global public speech.\footnote{Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017).} Forums are incidental to the primary

The same is true in examining the ability of homeowners’ associations or professional organizations to conscribe the speech of members or individuals.\footnote{See Claudia E. Haupt, Professional Speech, 125 YALE L.J. 1238, 1241–42 (2016).} The special purposes of universities, professional organizations, or homeowners’ associations — to confer knowledge, protect a professional identity, or create a distinct visual community — are distinct from the motives of online speech platforms. Moreover, the global scale and essential nature of private governance of online speech separate it in kind from the strictures governing individuals within these isolated organizations.

The law reasons by analogy, yet none of these analogies to private moderation of the public right of speech seem to precisely meet the descriptive nature of what online platforms are, or the normative results of what we want them to be. The following Part argues for a new kind of understanding: seeing these platforms’ regulation of speech as governance.

IV. THE NEW GOVERNORS

Thinking of online platforms from within the categories already established in First Amendment jurisprudence — as company towns, broadcasters, or editors — misses much of what is actually happening in these private spaces. Instead, analysis of online speech is best considered from the perspectives of private governance and self-regulation.\footnote{See, e.g., PASQUALE, supra note 125, at 140–68, 185–218 (arguing that terms of service or contracts are inappropriate or ineffective remedies in an essentially “feudal” sphere, id. at 144, and that platforms act as “sovereign[s]” over realms of life, id. at 163, 180); Freeman, supra note 15, at 636–64 (describing the ability of private firms to self-regulate in areas of public interest with and without government influence); Michael P. Vandenbergh, The Private Life of Public Law, 105 COLUM. L. REV. 2029, 2037–41 (2005) (discussing how private actors play an increasing role in the traditional government standard-setting, implementation, and enforcement functions through contracts and private agreements). On the role of voluntary self-regulation by private actors, see NEIL GUNNINGHAM ET AL., SMART REGULATION: DESIGNING ENVIRONMENTAL POLICY 167–70 (1998), which analyzes shortcomings of self-regulation, including lack of transparency and independent auditing, concern that performance is not being evaluated, and absence of real penalties for recalcitrants.}

Analyzing online platforms from the perspective of governance is both more descriptively accurate and more normatively useful in addressing the infrastructure of this ever-evolving private space. Platform governance does not fit neatly into any existing governance model, but
it does have features of existing governance models that support its categorization as governance. As Parts II and III demonstrated, platforms have a centralized body, an established set of laws or rules, ex ante and ex post procedures for adjudication of content against rules, and democratic values and culture; policies and rules are modified and updated through external input; platforms are economically subject to normative influence of citizen-users and are also collaborative with external networks like government and third-party groups. Another way to conceptualize the governance of online speech by platforms comes from administrative law, which has long implicated the motivations and systems created by private actors to self-regulate in ways that reflect the norms of a community. Perhaps most significantly, the idea of governance captures the power and scope these private platforms wield through their moderation systems and lends gravitas to their role in democratic culture. Changes in technology and the growth of the internet have resulted in a “revolution in the infrastructure of free expression.” The private platforms that created and control that infrastructure are the New Governors in the digital era.

How does this new concept of private platform governors normatively fit in our hopes and fears for the internet? For decades, legal scholars have moved between optimistic and pessimistic views of the future of online speech and long debated how — or whether — to constrain it. But the details of the private infrastructure of online speech were largely opaque. Does this new information and conception allay or augment scholarly concerns over the future of digital speech and democratic culture?

The realities of these platforms both underscore and relieve some of these fears. For the optimists, interviews with the architects of these

456 Balkin, supra note 11, at 2296.
457 Id.
458 Lessig was an early pessimist about the future of the internet, seeing it as a potential means of regulation and control. He specifically worried about the domination of the internet by commercial forces that could be manipulated and controlled by the state. Lessig, supra note 21, at 71. Boyle, Goldsmith, and Wu had similar concerns about the state co-opting private online intermediaries for enforcement. See Goldsmith & Wu, supra note 100; Boyle, supra note 100, at 202–04. In contrast, Balkin has been largely optimistic about the growth of the internet, the growth of platforms, and the ability of these new speech infrastructures to enhance the “possibility of democratic culture.” Balkin, supra note 7, at 46. But recently he too has become concerned about the future of online speech and democracy, arguing that private platforms and government can together regulate online speech with less transparency, disruption, and obtrusion than ever before. See Balkin, supra note 11, at 2342. Scholars like Citron, Norton, and Franks have instead long argued for working with private platforms to change their policies. See Bazelon, supra note 105, at 279–89; Citron, supra note 102, at 121–25; Citron & Norton, supra note 104, at 1488–84; Franks, supra note 103, at 681–88; cf. Citron & Franks, supra note 106, at 386–90 (discussing the need for governments to craft criminal statutes prohibiting the publication of revenge porn).
platform content-moderation systems show how the rules and procedures for moderating content are undergirded by American free speech norms and a democratic culture.\footnote{This is good news for Lessig, Balkin, and Benkler, given their concerns.} These ideas are also part of their corporate culture and sense of social responsibility. But perhaps more compellingly, platforms are economically responsive to the expectations and norms of their users. In order to achieve this responsiveness, they have developed an intricate system to both take down content their users don’t want to see and keep up as much content as possible. To do this has also meant they have often pushed back against government requests for takedown.\footnote{If this trend continues, it allays much of Balkin’s concern over collateral censorship in Old-School/New-School Speech Regulation. See Balkin, supra note 11.} Procedurally, platform content-moderation systems have many similarities to a legal system. Finally, platforms have a diverse pluralistic group of forces that informs updates of their content-moderation policies and procedures.

Not only is governance the descriptively correct way to understand platform content moderation, but it is also rhetorically and normatively correct. Historically, speech regulation has followed a dyadic model: a territorial government, with all the power that that invokes, has the boot on the neck of individual speakers or publishers.\footnote{Balkin, supra note 362 (manuscript at 4, 41).} The New Governors are part of a new model of free expression: a triadic model.\footnote{Id. (manuscript at 41–44). Balkin refers to this as a “pluralist” model, id. (manuscript at 4), and while that term is perhaps more accurate for the world of internet speech as a whole, for my focus here I prefer to use the term “triadic.”} In this new model, online speech platforms sit between the state and speakers and publishers. They have the role of empowering both individual speakers and publishers (as well as arguably minimizing the necessity of publishers to speaking and amplification), and their transnational private infrastructure tempers the power of the state to censor. These New Governors have profoundly equalized access to speech publication, centralized decentralized communities, opened vast new resources of communal knowledge, and created infinite ways to spread culture. Digital speech has created a global democratic culture,\footnote{Id. (manuscript at 41–44).} and the New Governors are the architects of the governance structure that runs it.

The system that these companies have put in place to match the expectations of users and to self-regulate is impressively intricate and responsive. But this system also presents some unquestionable downsides that grow increasingly apparent. These can be seen in two main concerns: (1) worries over loss of equal access to and participation in speech on these platforms; and correspondingly (2) lack of direct platform accountability to their users.
A. Equal Access

There is very little transparency from these private platforms, making it hard to accurately assess the extent to which we should be concerned about speech regulation, censorship, and collateral censorship. But separate from the question of secret government interference or collusion, private platforms are increasingly making their own choices around content moderation that give preferential treatment to some users over others. The threat of special rules for public figures or newsworthy events crystallizes the main value we need protected within this private governance structure in order to maintain a democratic culture: fair opportunity to participate.

In some ways, an ideal solution would be for these platforms to put their intricate systems of self-regulation to work to solve this problem themselves without regulatory interference. But the lack of an appeals system for individual users and the open acknowledgment of different treatment and rule sets for powerful users over others reveal that a fair opportunity to participate is not currently a prioritized part of platform moderation systems. In a limited sense, these problems are nothing new — they are quite similar to the concerns to democracy posed by a mass media captured by a powerful, wealthy elite. Before the internet, these concerns were addressed by imposing government regulation on mass media companies to ensure free speech and a healthy democracy. But unlike mass media, which was always in the hands of an exclusive few, the internet has been a force for free speech and democratic participation since its inception.

The internet has also made speech less expensive, more accessible, more generative, and more interactive than it had arguably ever been before. These aspects of online speech have led to the promotion and development of democratic culture, writes Balkin, “a form of social life in which unjust barriers of rank

464 These are the concerns expressed by Balkin, Lessig, Tushnet, and Wu. See Balkin, supra note 11, at 2308–14; Lessig, supra note 21, at 327–29; Tushnet, supra note 263, at 1002–15; Wu, supra note 61, at 317–18.


466 It is important to note that the uses of “public figure” and “newsworthiness” here differ from their meanings in the sense of communications or privacy torts.

467 Balkin, supra note 7, at 30.

468 Id. at 31.

469 See BENKLER, supra note 98; Lessig, supra note 21; Balkin, supra note 7, at 3–6.
and privilege are dissolved, and in which ordinary people gain a greater say over the institutions and practices that shape them and their futures. What makes a culture democratic, then, is not democratic governance, but democratic participation.”

Equal access to platforms is thus both an effect of a self-regulated and open internet and the cause of it, making regulation of this issue particularly difficult and paradoxical. Legislating one user rule set for all not only seems logistically problematic, but it would also likely reduce platforms’ incentives to moderate well. Such legislation, if constitutionally valid, would certainly run into many of the concerns raised by those who fear any regulation that might curb the robust power of §230 immunity. This is why any proposed regulation — be it entirely new laws or modest changes to §230 — should look carefully at how and why the New Governors actually moderate speech. Such, if any, regulation should work with an understanding of the intricate self-regulatory structure already in place in order to be the most effective for users.

B. Accountability

Even without issues of equal access to participation, the central difficulty in simply allowing these systems to self-regulate in a way that takes into account the values and rights of their users is that it leaves users essentially powerless. There is no longer any illusion about the scope and impact of private companies in online platforms and speech. These platforms are beholden to their corporate values, to the foundational norms of American free speech, and to creating a platform where users will want to engage. Only the last of these three motivations for moderating content gives the user any “power,” and then only in an indirect and amorphous way.

Moreover, while it initially seems like a positive source of accountability that these systems are indirectly democratically responsive to users’ norms, it also creates inherently undemocratic consequences. At first, adaptability appears to be a positive attribute of the system: its ability to rapidly adapt its rules and code to reflect the norms and values of users. But that feature has two bugs: in order to engage with the most users, a platform is (1) disincentivized to allow antinormative content, and (2) incentivized to create perfect filtering to show a user only content

470 Balkin, supra note 7, at 35.
472 This was a central concern of Lessig’s — that the internet would be captured by large corporations. See generally LESSIG, supra note 21.
473 For an excellent discussion of this interplay between corporate power, inequitable markets, and democratic capacity of citizens and users, see generally K. SABEEL RAHMAN, DEMOCRACY AGAINST DOMINATION (2017).
that meets her tastes. These problems are interchangeably known as the so-called echo-chamber effect, which creates an antidemocratic space in which people are shown things with which they already associate and agree, leading to nondeliberative polarization. “It has never been our ideal — constitutionally at least — for democracy to be a perfect reflection of the present temperature of the people.”

Whether through algorithmic filtering or new content rules, as platforms regress to the normative mean, users will not only be exposed to less diverse content, but they will also be less able to post antinormative content as external and internal content-moderation policies standardize across platforms.

Since the 2016 American presidential election, the lack of accountability of these sites to their users and to the government in policing fake news, commercial speech, or political speech has come to the fore of public consciousness. In statements directly following the election of Trump as President, Zuckerberg emphatically denied the role of fake news in the result. But due to many of the factors discussed here — media pressure, corporate responsibility, and user expectations — Facebook was forced to start tackling the issue. Yet the power of these new threats to “spread[] so quickly and persuade[] so effectively” might make these indirect systems of accountability unexpectedly slow

---

474 LESSIG, supra note 21, at 331.

476 Most notably, in late 2017 it was revealed that hundreds of thousands of dollars in ads placed on Facebook during the election had actually come from Russia-linked groups. See Mike Isaac & Scott Shane, Facebook’s Russia-Linked Ads Came in Many Disguises, N.Y. TIMES (Oct. 2, 2017), http://nyti.ms/2zgeVIj [https://perma.cc/SE8S-X72P]; Carol D. Leonnig et al., Russian Firm Tied to Pro-Kremlin Propaganda Advertised on Facebook During Election, WASH. POST (Sept. 6, 2017), http://wapo.st/2C4pdoH [https://perma.cc/699Y-EP6V].

477 Following the Russia-linked ads, many platforms have been moving to police more heavily all ad content relating to important issues of political speech. See, e.g., Erik Schelzig, Twitter Shuts Down Blackburn Campaign Announcement Video, AP NEWS (Oct. 9, 2017), https://apnews.com/0d8828bd7d204a0f61172628d0a756 [https://perma.cc/U97N-37E5]; see also Carla Herreria, Mark Zuckerberg: “I Regret” Rejecting Idea that It Swayed Voters, THE GUARDIAN (Nov. 10, 2016, 10:01 PM), https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-us-election-mark-zuckerberg-donald-trump [https://perma.cc/PKD5-BHRW].


for dealing with such emerging threats and issues. It also makes clear that some insertion of traditional government agency functions — such as regulation of commercial speech — when matched with an accurate understanding of how these platforms currently moderate content, could provide a potential answer to such issues of accountability.

The lack of accountability is also troubling in that it lays bare our dependence on these private platforms to exercise our public rights. Besides exit or leveraging of government, media, or third-party lobbying groups, users are simply dependent on the whims of these corporations. While platforms are arguably also susceptible to the whims of their users, this is entirely indirect — through advertising views, not through any kind of direct market empowerment. One regulatory possibility might be a type of shareholder model — but this fails not only because Zuckerberg owns controlling shares of Facebook, but also because shareholder values of maximizing company profits are perhaps not well matched with user concerns over equal access and democratic accountability. One potential nonregulatory solution to this problem would be for these corporations to register as public benefit corporations, which would allow public benefit to be a charter purpose in addition to the traditional maximizing profit goal.

Another avenue would be for platforms to voluntarily take up a commitment to a notion of “technological due process.” In this groundbreaking model for best practices in agency use of technology, Citron advocates for a model that understands the trade-offs of “automation and human discretion,” protects individuals’ rights to notice and hearings, and gives transparency to rulemaking and adjudication. Of course, these private platforms have little motivation to surrender power as in a public benefit corporation, or to adopt the rules and transparency ideas of Citron’s technological due process requirements — but they

481 So far private nongovernmental groups have focused on this. For example, ProPublica has launched a browser attachment to help monitor political ads on online platforms. See Julia Angwin & Jeff Larson, Help Us Monitor Political Ads Online, PROPUBLICA (Sept. 7, 2017, 10:00 AM), https://www.propublica.org/article/help-us-monitor-political-ads-online [https://perma.cc/A35R-WHHR]. For an excellent and complete discussion of how potential regulation or change should take into account the realities of platforms and moderation, see Syed, supra note 480.
484 Id. at 1301.
might if they fear the alternative would result in more restrictive regulation. Should these platforms come under agency regulation, however, the concerns detailed by Citron’s notion of technological due process combined with an accurate understanding of how such companies self-regulate will be essential to crafting responsive and accurate oversight.

CONCLUSION

As the Facebook Live video of Philando Castile’s death demonstrates, content published on platforms implicates social policy, law, culture, and the world. Yet, despite the essential nature of these platforms to modern free speech and democratic culture, very little is known about how or why the platforms curate user content. This Article set out to answer these questions. It began with an overview of the legal framework behind private platforms’ broad immunity to moderate content. This framework comes from §230, the purposes of which were both to encourage platforms to be Good Samaritans by taking an active role in removing offensive content and to protect users’ rights by avoiding free speech problems of collateral censorship. With this background, this Article explored why platforms moderate despite the broad immunity of §230. Through interviews with former platform architects and archived materials, this Article argued that platforms moderate content partly because of American free speech norms and corporate responsibility, but most importantly, because of the economic necessity of creating an environment that reflects the expectations of their users.

Beyond §230, courts have struggled with how to conceptualize online platforms within First Amendment doctrine: as company towns, as broadcasters, or as editors. This Article has argued that the answer to how best to conceptualize platforms lies outside current categories in First Amendment doctrine. Through internal documents, archived materials, interviews with platform executives, and conversations with content moderators, this Article showed that platforms have developed a system of governance, with a detailed list of rules, trained human decisionmaking to apply those rules, and reliance on a system of external influence to update and amend those rules. Platforms are the New Governors of online speech. These New Governors are private self-regulating entities that are economically and normatively motivated to

---

485 The window for using governmental threat to produce a voluntary result might be closing as the scope and power of these companies make them increasingly difficult to regulate. See, for example, Google’s lengthy and robust attempts to push back at the European Court of Justice judgment mandating the “Right to Be Forgotten.” The Right to Be Forgotten (Google v. Spain), ELECTRONIC PRIVACY INFO. CTR., https://epic.org/privacy/right-to-be-forgotten/ [https://perma.cc/GjXT-AWR4].

486 While Castile’s live-streamed death crystallized the conversation around police brutality and racism in America, it is necessary to note that the officer who shot him was ultimately acquitted. See Mitch Smith, Minnesota Officer Acquitted in Killing of Philando Castile, N.Y. TIMES (June 16, 2017), http://nyti.ms/2CrMkJF [https://perma.cc/8ETE-LLZE].
reflect the democratic culture and free speech expectations of their users. But these incentives might no longer be enough.

The impact of the video of Philando Castile, the public outcry over Napalm Girl, the alarm expressed at the Zuckerberg Town Hall meeting, and the separate Twitter Rules for President Trump all reflect a central concern: a need for equal access to participation and more direct platform accountability to users. These New Governors play an essential new role in freedom of expression. The platforms are the products of a self-regulated and open internet, but they are only as democratic as the democratic culture and democratic participation reflected in them. Any proposed regulation — be it entirely new laws or modest changes to § 230 — should look carefully at how and why the New Governors actually moderate speech. Such, if any, regulation should work with an understanding of the intricate self-regulatory structure already in place in order to be the most effective for users and preserve the democratizing power of online platforms.