

Christine M. Korsgaard is widely regarded as one of the preeminent proponents of a Kantian moral philosophy. This reputation is likely to be enhanced by the original, provocative, and carefully developed arguments in these two volumes. Self-Constitution (hereafter SC) is a sustained defense of the thesis that, by acting, human beings constitute themselves as agents and thereby commit themselves to principles of practical reason that include a significant moral principle, namely, Kant’s categorical imperative. Many of the essays in The Constitution of Agency (hereafter CA) also explain parts of the argument for this thesis, and virtually all the other essays are on related topics. In this respect, these are companion books (SC is based on Korsgaard’s 2002 Locke Lectures; all but two of the ten essays in CA have appeared elsewhere, in most cases from 1996 to 1999). These works also restate and develop much of Korsgaard’s position in her earlier book, The Sources of Normativity (Cambridge UK: Cambridge University Press, 1996; hereafter SN), but with greater emphasis on agency. Though Korsgaard writes as an avowed Kantian, she maintains that in important respects her normative outlook, and Kant’s, are shared by Plato and Aristotle. Familiar divisions between the ancients and the moderns are replaced, in Korsgaard’s striking interpretation, with a division between, on the one hand, Plato, Aristotle and Kant—“the heroes of this book” (SC, p. 27)—and, on the other hand, both empiricists, especially Hume, and an array of rationalists such as Samuel Clarke, Richard Price, W. D. Ross, and H. A. Prichard. Korsgaard moves seamlessly between claims of historical scholarship and arguments for philosophical positions in a contemporary context. These volumes are likely to be controversial on both fronts.

In the first section I will summarize Korsgaard’s main line of argument in SC and related essays in CA, and in the second I will then raise some questions about the persuasiveness of this argument in section II.

I

Korsgaard begins with the claim that we human beings, in contrast to other animals, are aware of the grounds of our beliefs and actions. As a result, we have the ability and need to control our beliefs and actions. We have no choice, Korsgaard says, but to decide what to think and do on the basis of reason—that is, on the basis of normative principles or laws—though we may do this well or badly. This is the sense in which we are rational animals. That we are rational in this sense is a fundamental assumption of Korsgaard’s overall argument.
No consideration is given to possible skeptical challenges to this assumption, but many will agree that it is a plausible contention. Korsgaard initially describes this feature of human nature—that we have to decide on the basis of reason—as a curse, as our plight, as something to which we are condemned. Perhaps these are simply rhetorical flourishes, but they make it surprising that she eventually concludes that this is also the source of the value of our humanity, something we ought to respect. Though Korsgaard regularly suggests that there are important parallels between action and belief, most of her arguments concern action. A distinctive feature of her account is that what is given to us by human nature is not the fact of being an agent, but the task of becoming an agent. In acting we continually constitute ourselves as agents, as the authors of what we do. “[T]he function of action is self-constitution” (SC, p. xii). Hence, becoming an agent is the work or job demanded of us by our nature, and it is something we may reasonably hold one another responsible for performing well (see SC, pp. 26, 175). Performing it well requires that we achieve a form of unity, the unity that constitutes being an agent, and we achieve this unity to the extent that we are guided by the principles of practical reason, the principles that define what it means to act. If we are not trying to conform to these principles, Korsgaard says, we are not acting at all. Since we need to act, the normativity of these principles cannot be challenged.

Korsgaard’s argument is based on what she calls “the metaphysics of normativity,” a theory she claims was accepted by Plato and Kant, but was best explained by Aristotle. According to this theory, the difference between being a “mere heap” and being an object of a certain kind is that the latter, but not the former, has a function. For example, the function of a house is to provide shelter: to know this is to know what a house is. This functional account implies that objects have “constitutive standards.” These are “standards that apply to a thing simply in virtue of its being the kind of thing that it is” (SC, p. 28). For instance, in order for something to be a good house, it must have a roof that protects us from the rain. Something similar can be said about activities such as swimming: these are defined by their function and their function gives rise to “constitutive principles.” For example, the function of swimming is to stay afloat, and so a person is swimming well only insofar as the movement of his or her limbs contributes to this aim. Moreover, if a person in the water with limbs in motion is not trying to conform to this principle, the person cannot be said to be swimming at all, but merely to be splashing around. Hence, constitutive principles are both descriptive and normative: they tell us what an activity is and what it means to perform it well. For this reason, the normative authority of these principles is secure: it would not make sense to ask why a good swimmer needs to stay afloat. Of course, it is possible to try to conform to a constitutive principle and fail. But swimming poorly is not a different activity than swimming well: it is the same activity, only it is badly done.

The metaphysics of normativity is also applied to living things such as animals. Their function is simply to maintain and reproduce their form. The function of a giraffe is to make itself into a giraffe (more on this odd locution below) as well as to produce other giraffes. Korsgaard grants that this is a teleological understanding, based on Aristotle, but she denies that it is inconsistent with Darwinian biology. Her claim is just that the purpose of things is to function properly, not that they have some natural purpose beyond this. She argues that identifying things by their functions is, at least initially, the only way we can conceptualize the world: as agents, we need to see the world in terms of things that do or produce something, and hence as things that may be tools or obstacles in fulfilling our purposes. Science, it seems, does not undermine this (but see SC, pp. 41, 114).

As we have seen, a primary function of human beings is to act, and this requires unifying ourselves as agents by following principles of practical reason. Korsgaard says that we need
three such principles: an instrumental principle, a moral principle, and a principle of prudence or self-interest. The instrumental principle is that we should take the means necessary to achieve our ends. This is (what Korsgaard calls) Kant’s “hypothetical imperative.” The moral principle is Kant’s categorical imperative (though, as will be evident shortly, Korsgaard uses this phrase in more than one way). However, these two principles are not sufficient because they do not tell us how to balance our various ends. (I would add that they also do not tell us what all our ends should be, an issue I will return to below.) This is why we need a prudential principle. A familiar candidate is that our ultimate prudential end should be to maximize the satisfaction of our desires. But Korsgaard rejects arguments for this as dogmatic assertions. Moreover, she professes not to know what the prudential principle is. All she says is that this “missing principle” must be a formal principle that unifies our agency. In any case, she maintains that her instrumental and moral principles are formal principles that unify our agency, and almost all of her attention is devoted to them.

Korsgaard argues at some length that neither empiricism nor rationalism can establish the normativity of the principles of practical reason (see in particular SC, chap. 4 and CA, chap. 1). Her argument focuses especially on the instrumental principle. In brief, empiricism can explain why this principle motivates us, but it cannot explain why it is binding, while rationalism (or realism) simply declares that it is binding, but cannot explain why it should motivate us. Korsgaard claims that her constitutivist position succeeds where these accounts fail. According to Korsgaard, insofar as you are striving to be an agent by willing and acting, you are determining yourself to be the cause of some end. There are two essential features of this. The first is efficacy: being the cause of some end. An agent is bound by the hypothetical imperative because the only way to cause an end is to take the means to the end. The second feature is autonomy: determining yourself to be that cause. An agent is bound by the categorical imperative because the only way to determine yourself to be the cause is to act on a maxim that could be willed as a universal law. More will need to be said about these claims, but this is the basic structure of the argument. In sum: insofar as a person is acting, a person is committed to being efficacious and autonomous, and insofar as a person is so committed, a person is committed to the hypothetical and categorical imperatives. These imperatives are binding because they are constitutive of action: there is no way to act except by trying to follow these principles. And we are motivated to follow these principles because the human situation is such that we have no choice but to act. Hence, constitutivism can do what neither empiricism nor rationalism could do.

The hypothetical imperative is the more obvious of the two principles: it is hard to see how I could be committed to an end unless I were also committed to taking the means necessary to achieve it. If I give in to each temptation to avoid the means, then I am not really willing the end. “Conformity to the hypothetical imperative is thus constitutive of having a will” (SC, p. 70). Moreover, since willing is necessary to being an agent, following this principle is also constitutive of being an agent. Insofar as I give in to the temptations, I am not really an agent at all.

Korsgaard has a related argument for the categorical imperative. She claims that the only alternative to acting on a maxim that could be willed as a universal law is what she calls “particularist willing”: acting on a maxim that applies to only one case. According to Korsgaard, this is not willing at all. In deliberation, we face an array of incentives to act in various ways. Successful deliberation culminates in willing and acting on the basis of a choice to follow one of these incentives, and this requires that there be something in addition to the incentives—you—that is identified by the principle of choice on which you act. “You regard the choice as yours . . . because you regard the principle of choice as expressive, or representative, of
yourself—of your own causality” (SC, p. 75). Acting on the basis of your principle is the only way to determine yourself to be the cause of something. But in particularistic willing there is no distinction between you and the incentives: you simply identify with one of the incentives, not as an instance of a type, but “in its full particularity,” and so there is no you, no principle of choice, in addition to your incentives. “If you have a particularistic will,” Korsgaard says, “you are not one person, but a series, a mere heap, of unrelated impulses” (SC, p. 76). In this case, you are not causing anything; an incentive within you is simply bringing something about. Hence, particularistic willing is not possible. If you are willing at all, you are trying to act on a maxim that could be willed as a universal law. Following the categorical imperative, so understood, is another constitutive feature of willing and thus of being an agent (for earlier versions of this argument, see SN, pp. 225–33 and CA, pp. 120–24).

There are several features of the categorical imperative that this argument is supposed to establish that should be noted. First, Korsgaard thinks that the maxim that is to be willed as a universal law must refer, not simply to an act, but to “an-act-for-the-sake-of-an-end” (SC, p. 70) that is provisional (we may discover a reason to revise it later) and may have any degree of specificity. The implication is that our maxims, though they must be universal in the sense of referring to some type, should not be general rules (she thinks that Kant believed this as well; see SC, p. 15). Second, because the maxim refers to an-act-for-the-sake-of-an-end, Korsgaard thinks that the hypothetical imperative is not really separate from the categorical imperative, but an aspect of it that specifies that the law that is needed is a causal law (see SC, pp. 70–72, 81). Nonetheless, she usually speaks as if these two imperatives are distinct because they draw attention to the aforementioned distinction between efficacy and autonomy. Finally, the categorical imperative purportedly established by the argument against particularistic willing only says that one should not act on maxims unless one could will that they be universal laws for oneself. This is not sufficient to establish the moral law (the categorical imperative as Kant understood it), which requires us to act on maxims that all rational beings could act on together. In order to establish the moral law, Korsgaard says, it would also have to be shown that our maxims have to be willed as universal laws for all rational beings and that these laws are public in the sense of having normative force for all rational beings (for the distinction between these two senses of the categorical imperative, see SC, p. 80 and SN, pp. 98–100). So how do we do this?

Korsgaard’s answer to this question is based on an extended analysis in which she aligns her position with the account of justice in Plato’s Republic, in particular with the argument based on the analogy of the three parts of the city and the three parts of the soul (see SC, chchap. 6–9 and CA, chap. 3). Socrates says that “those who are all bad and completely unjust are completely incapable of accomplishing anything” (Rep. 352c, as quoted in SC, p. 136). Korsgaard agrees: insofar as we are really acting—are truly efficacious and autonomous—we are morally good.

In Korsgaard’s analysis of human action, we are presented with what Kant calls incentives (representations of things as in some way attractive or repellent) and we need principles in order to decide which of these incentives to follow. Something similar is true of all animals, but in non-human animals the principles take the form of instincts, whereas in human animals the principles need to be chosen on the basis of reason, that is, on the basis of the hypothetical and categorical imperatives. As rational beings, we are liberated from being determined to act from instincts. This means we have to choose how to live, a phenomenon Korsgaard calls “our expulsion from the Garden” and our “essential homelessness” (SC, p. 118), that to which she said earlier we are condemned. Moreover, as rational beings, we have responses to our incentives in the form of what Kant calls inclinations (for example,
we want the thing that appears attractive). This means that our souls have parts: at a minimum, there is the part that contains inclinations and the part that decides how to respond to these inclinations. Since the soul has these parts, Korsgaard thinks, they must be unified before we can act (in contrast to non-human animals, where unification is a natural state, not a task to be achieved). This is the function of rational deliberation. Unification implies making choices based on principles that constitute one’s own individual form or identity as a particular person (one’s practical identity, in the terminology Korsgaard introduced in SN). But unification is also supposed to have moral implications.

The connection with Plato’s argument in the Republic is now more evident. In drawing this connection, Korsgaard contrasts what she calls Plato’s “Constitutional Model” of the soul with the “Combat Model” that she associates with (but does not quite attribute to) Hume. According to the Combat Model, the soul has two forces, reason and passion, that combat one another, and the virtuous person follows reason. Korsgaard thinks that this model does not provide an intelligible account of what it means to be an agent: the agent would have to be something over and above these two forces that chooses between them (it cannot simply be the result of the combat), but it is hard to see what this something could be if it is neither passion nor reason. She thinks a much better account of agency is provided by Plato’s Constitutional Model. According to this model, just as a group of individual persons needs to be unified in a certain way in order to act as a city, as a collective agent, so the parts of a person’s soul must be unified in a certain way in order to act as an individual agent. In both cases, the needed unity is established by a constitution that assigns a role to each of the parts. In the case of the city, the people propose a course of action, the rulers decide whether or not to accept the proposal, and the auxiliaries carry out the decision. In the case of the individual person, appetite proposes an action, reason decides whether or not to take the action, and spirit carries out the decision. Since the constitution determines the roles played by each of the parts, collective agency and individual agency are constituted by the respective parts following their assigned roles (in both cases, she says, the parts would be a mere heap without the constitution). Since justice is defined as following these roles and not interfering with the roles of the other parts (see Rep. 433a–b), justice is needed for agency. Socrates is correct.

Korsgaard argues that Plato is thereby committed to a procedural rather than a substantive conception of justice: as long as each part plays its proper role (the correct procedure), the result is just. The procedure is warranted, not because it always or even typically produces what we independently know to be substantively just results, but because following the procedure is the only way to act. Hence, the procedure confers normativity on the results, not the other way around. That which we regard as most substantively just can be nothing else than that which the most ideal procedure would produce; and for Korsgaard this is the procedure that best unifies agency. She thinks this is the case for both the city and the soul.

Korsgaard also maintains that Kant accepted the Constitutional Model and not the Combat Model. With respect to the individual person, inclination proposes doing some act A for some end E (for example, pleasure). Reason decides whether or not to accept the proposal by asking whether it can will doing A for E as a universal law. “The decision, she says, “takes the form of a legislative act” (SC, p. 154). Korsgaard grants that Kant has little to say about the spirited part of the soul, but she finds some resources in a few remarks by Kant on courage. With respect to the city (in Kant’s political philosophy), the state must have a republican constitution in which legislative authority ultimately belongs to “the general will of its people to the reciprocal enforcement of rights” (SC, p. 155). There are questions about what is required for there to be such a constitution. But the implication, I take it, is
that only insofar as there is a constitution does the state truly act, and insofar as it does so act, it acts justly.

There is an obvious problem for the Constitutional Model: if a soul or city truly acts only insofar as it acts justly or well, then it might seem that it is not possible to act unjustly or badly. Korsgaard thinks that this problem arises for both Plato and Kant, and she thinks that their responses to the problem are substantially the same. Bad action is action because it is governed by some principle and so the agent is to some extent unified. But bad action is bad because it is not governed by the principle of reason, the principle assigned by the constitution to rule, and so the agent is not fully unified in the way that is necessary for genuine efficacious and autonomous action. Korsgaard thinks that it is possible to act on a principle that is more or less close to the principle of reason. In this sense, there are degrees of action that correspond to degrees of the badness of action. She believes that Plato has the most detailed analysis of this, and she largely endorses his account of the timocratic, oligarchic, democratic and tyrannical souls (in Books 8 and 9 of the Republic). In each case, a principle is followed, but it is increasingly remote from the principle of reason and so from the aim of acting for the good of the soul as a whole. The tyrannical soul, she thinks, is barely acting at all because it has made itself into a “force of nature.” She is not sure whether or not the tyrannical soul can be held responsible for this stance, but in all the other cases the persons have the same degree of responsibility for their failure to achieve full agency because their job as human beings is to become an agent in the full sense.

Only after the extended argument that Plato and Kant share the Constitutional Model does Korsgaard directly defend the thesis that, insofar as we are agents, we are committed to morality. In her words, “respect for humanity is a necessary condition of effective action” (SC, p. 206). Korsgaard’s defense of this thesis in chapter 9 of SC is long and multifarious; in contrast to the rest of the book, it is not always easy to discern the main line of argument. It is clear, however, that the crucial step involves a distinction between private reasons and public reasons. Both kinds of reasons are subject to a universalizability requirement, the two senses of the categorical imperative discussed earlier, and the key contention is that the two requirements go together: “legislating for oneself, and legislating for the Kingdom of Ends, are one and the same thing” (SC, p. 206).

A private reason is the possession of one person. For example, the fact that obtaining X will make me happy may give me a private reason to obtain X. That I have this private reason does not by itself give you a reason to obtain X for me. However, my having this reason commits me to thinking that the fact that obtaining Y makes you happy gives you a private reason to obtain Y. This is the implication of the universalizability requirement for private reasons (on the face of it, this requirement goes beyond what the argument against particularistic willing was meant to establish directly). Now, suppose all reasons were merely private. This would leave “us each with our own system of private reasons, which don’t have to be consistent with anyone else’s” (SC, p. 191). For example, if X and Y were the very same object, then I would have a private reason to obtain the object and so would you. We would be in a practical conflict. Moreover, Korsgaard thinks, if all reasons were merely private, we would be at war with one another: each of us might have a reason to shoot the other in order to obtain the object, and each of us would have to acknowledge that the other has this reason. But there is nothing more to be said, and so each of us would have a reason to shoot first.

Public reasons are different. Their “normative force can extend across the boundaries between people” (SC, p. 191). Reverting to the previous example, if I had a public reason to shoot you, then “I must be able to will that you should shoot me” (SC, p. 192). This is
the implication of the universalizability requirement for public reasons. Since I cannot will this, I cannot suppose that I have a public reason to shoot you. Hence, from the standpoint of public reasons, neither of us has a reason to shoot the other. The universality requirement of public reasons, Korsgaard says, brings us “into moral territory.”

Korsgaard maintains that public reasons are necessary for shared deliberation and for important kinds of personal interaction such as friendship and marriage (as well as more ordinary interactions such as making promises). She argues at some length that public reasons are possible because through shared deliberation persons are able to establish “unified wills” in which each party regards the reasons of the other as having normative force, not just for the other, but for oneself (all of which she says, with little explanation, requires meeting in the noumenal rather than the empirical world). But what Korsgaard needs to establish is that there is an important sense in which public reasons are not only possible but necessary, and it is not easy to make out her basis for thinking this.

Korsgaard herself raises an obvious objection: until the person with merely private reasons (the private reasoner) decides to interact with other persons in the way just described, “he has no reason to accord normative force to the reasons of other people” (SC, p. 200). In the initial part of Korsgaard’s response to this objection, she appears to argue that a private reasoner would not be logically required to endorse public reasons (see SC, pp. 200–02; cf. SN, pp. 132–34). At any rate, she concludes that a private reasoner cannot choose or decide to regard the reasons of other persons as having normative force. But this does not mean that we are all locked “in our own little system of private reasons.” This is because “responding to another’s reasons as normative is the default position” (SC, p. 202, emphasis added).

We have to “work to ignore someone else’s reasons” because it just is the case that “reasons are public” (SC, p. 202; cf. SN, pp. 134–36). Presumably this means that no one has merely private reasons: if a person is engaged in practical reason at all, the person is committed to endorsing public reasons. This is the most important claim in Korsgaard’s argument for the thesis that being an agent commits a person to morality. But what does this claim mean and why does she think it is true?

The phrase “default position” might be understood as the empirical claim that, as a matter of fact, people just do regard one another’s reasons as having normative force. However, the most that could plausibly be maintained on empirical grounds is that most people regard the reasons of some other people as having normative force. Any such qualified empirical fact would obviously not be sufficient to establish the rational necessity of respecting humanity as such. What Korsgaard needs to establish is that, if an agent is logically committed to the first form of the categorical imperative (on account of the argument against particularistic willing), then that agent is also logically committed—and not merely, as a matter of fact, psychologically disposed—to regarding (at least some of) the reasons of all other persons as having normative force. She clearly thinks that this is true. However, she does not make it obvious why she thinks this. Though Korsgaard does not put it in these terms, it might be thought that her position is reminiscent of Kant’s claim that the moral law is a “fact of pure reason” that cannot be proved (see Critique of Practical Reason, 5:47). However, if this is so, this is a fact that requires some elucidation.

In SN, Korsgaard defended the claim that reasons are essentially public by appealing to our social nature and considerations analogous to Wittgenstein’s private language argument (see SN, pp. 135–45). She does not repeat or develop those claims here. But she does borrow another line of argument from SN. Korsgaard maintains that a person with a particularistic will would lack self-respect and that a person with (what may be called) a universalistic will—someone who is committed to acting on a maxim that could be willed as a universal
law (the first form of the categorical imperative)—not only has self-respect, but is also committed to respecting the humanity in his own person. The idea is that the person with a particularistic will would be a mere heap of incentives: there would be no self to respect and no commitment to respecting humanity. By contrast, the person with a universalistic will has constituted a self that is to be respected and thereby presupposes respect for the humanity that makes constitution of this self both possible and necessary. Korsgaard notes that it might be asked whether a person could respect the humanity in his own person, but not that in other persons (just it was asked whether a person could have private but no public reasons). In response, she says that this question is “ill-formed” because what is individually a person’s own is not his humanity, but what he makes of it, his practical identity. That identity, she says, “depends on respect for humanity in general” (SC, p. 204). This is an argument Korsgaard has long championed (for example, see SN, pp. 120–24, 129–30), now put into service in her self-constitution account of agency. In effect, her claim is that the presuppositions of the argument against particularistic willing, properly understood, commit us to what from a Kantian standpoint is the heart of morality: respect for humanity as such.

II

Korsgaard’s historical claims, in particular her attempts to align Kant with Plato and Aristotle, are strikingly original and forcefully argued. But they are perplexing in a number of respects. For example, it is not at all obvious that in the Republic Plato is committed to a procedural account of justice similar to Kant’s and that this can be understood without any reference to the form of the good (as Korsgaard evidently thinks). For both Plato and Kant, reason should govern the soul, but there is no basis for thinking that in Plato reason consists of applying the categorical imperative. Again, Korsgaard may be correct in supposing that, for Aristotle, a virtuous action must pass a universalizability test that applies to a detailed specification of the action (see SC, p. 17), but this is very different than saying, in a manner similar to Kant, that this test is sufficient for determining virtuous action. Korsgaard seems to grant this specific point, but this admission is the tip of the iceberg with respect to the general issue about the extent to which Kantian formalism can be reconciled with Aristotelian eudaimonism. Leaving aside claims of historical scholarship, as I will mostly do here, there are philosophical perplexities concerning Korsgaard’s argument that raise a question about whether or not her proceduralism can avoid appeals to the normativity of nature that are central to the Aristotelian tradition.

In many respects, her argument that being an agent has significant normative commitments is a compelling one. For example, she argues persuasively that being an agent requires trying to follow the hypothetical imperative similar to Kant’s and that this can be understood without any reference to the form of the good (as Korsgaard evidently thinks). For both Plato and Kant, reason should govern the soul, but there is no basis for thinking that in Plato reason consists of applying the categorical imperative. Again, Korsgaard may be correct in supposing that, for Aristotle, a virtuous action must pass a universalizability test that applies to a detailed specification of the action (see SC, p. 17), but this is very different than saying, in a manner similar to Kant, that this test is sufficient for determining virtuous action. Korsgaard seems to grant this specific point, but this admission is the tip of the iceberg with respect to the general issue about the extent to which Kantian formalism can be reconciled with Aristotelian eudaimonism. Leaving aside claims of historical scholarship, as I will mostly do here, there are philosophical perplexities concerning Korsgaard’s argument that raise a question about whether or not her proceduralism can avoid appeals to the normativity of nature that are central to the Aristotelian tradition.

In many respects, her argument that being an agent has significant normative commitments is a compelling one. For example, she argues persuasively that being an agent requires trying to follow the hypothetical imperative similar to Kant’s and precludes particularistic willing. But the last step, that one of these commitments is morality, in the form of respecting the humanity in all persons, is not compelling. As presented by Korsgaard, nature makes possible and necessitates that we choose on the basis of reasons. Let us call this the dictate of rational nature. She thinks that the only way to fulfill this dictate is to have a universalistic rather than a particularistic will.

Suppose we agree with this. How does respect for humanity come into the picture? There are many things it might be thought that nature dictates—for example, that we eat and sleep. No doubt we ought to accept these things as necessities of life. But more is needed to show that we ought to respect the nature that renders eating and sleeping necessary for life, or that in striving to fulfill this dictate we are committed to respecting the nature that produces it.

Korsgaard thinks that acting requires not only having a universalistic will, but valuing the specific practical identity that is constituted by this will. And we cannot value our
practical identity, she thinks, unless we value the rational human identity that dictates that we have a practical identity: “So in valuing ourselves as the bearers of contingent practical identities . . . we are also valuing ourselves as rational beings. For by doing that we are endorsing a reason that arises from our rational nature—namely, our need to have reasons” (SC, p. 25; cf. pp. 211–12). However, Korsgaard believes that we have no choice in this: we must choose and act, we must have reasons in order to do this, and we must have practical identities in order to have these reasons (see SC, pp. 1, 23–24). “Carving out a personal identity for which we are responsible,” she says, “is one of the inescapable tasks of human life” (SC, p. 24). But if we have no choice but to heed the dictate of our rational nature by embracing a personal identity, then the fact that we embrace some identity can hardly be said to imply that we endorse the value of the rational nature that requires this. Consider an analogy: if prisoners are herded into a room to hear either the warden or the chief guard speak, then the fact that they can choose which speech to hear does not commit them to endorsing the value of the prison system that requires them to hear one of them. In general, if we do something because we are compelled to do it, we are not thereby endorsing the value of what compels us to do it. In both of these cases, there would be reason to endorse the value of what compels us—human nature and the prison system respectively—if these had normative authority for us. But compliance by itself does not commit a person to the belief that they have this authority.

It might be said that there is an alternative to willing and hence to respecting humanity: we could choose to remain a heap of impulses and so not to be an agent at all. In that case, there is no commitment to respecting humanity. But since virtually all of us choose to will and act, we are committed to respecting humanity. In deciding to fulfill the dictate of rational nature we have implicitly endorsed the value of that nature.

But is this really so? In this scenario, there is presumably some sense in which we have decided that willing and acting are better than remaining a heap of impulses. But this decision falls short of unconditionally endorsing the value of our rational nature. Return to the prisoner analogy. It may be said that there is an alternative to hearing the speeches: a prisoner can resist going into the room. However, the prisoner who does not resist cannot be said thereby to be endorsing the value of being compelled to hear the speeches. He may simply be making the best of a bad situation. Moreover, even if we endorse the value of our practical identities (as Korsgaard supposes), it is not clear why doing so requires us to endorse the value of the nature that dictates that we have such identities or else remain a heap of impulses. We might prefer to be animals that act on instinct and hence are not compelled to choose between these two alternatives. In general, it is quite intelligible to endorse something that was chosen in circumstances that we did not choose without endorsing the value of those circumstances. We might simply begrudgingly go along with the circumstances. For example, a person might come to value a career that is much less than she had originally hoped for, but was dictated by the circumstances of her life. By choosing that career as the best she could do under the circumstances, she has not thereby committed herself to the value of those circumstances. She is just doing the best she can under the circumstances. Korsgaard has not shown that choosing to will and act rather than remain a heap of impulses commits a person to respecting the nature that dictates this choice. And in any case, she says that action is necessary (for example, see SC, pp. 1–2, 32, 87). Though I might end up a heap of impulses, I cannot choose to be a heap. For once I choose I am already an agent.

In effect, Korsgaard thinks that as rational creatures we are committed to respecting rationality wherever it is found. No doubt there are reasons to respect rationality, but it is not obvious that being a rational creature all by itself provides such a reason (though it
may enable us to ascertain the reason). It might be said that it would be inconsistent to act rationally but fail to respect rationality. But if we have no real choice but to do this (that is, to act on the basis of reasons, which may or may not be good reasons) because this is the plight we are compelled to accept, it is difficult to see the force of this argument. What needs to be shown is that the dictate of rational nature has normative authority for us.

There is another aspect of this issue. Korsgaard says that “all values and reasons are human creations” in which form (universal law) is imposed on some matter such as a desire or interest (SC, p. 209; cf. p. 123). This is an expression of her proceduralism: values and reasons are created by following the procedure of formulating maxims that could be universal laws. But the most fundamental value for Korsgaard is that we must fulfill the task assigned to us by nature: “nature sets each human being a task,” she says, and this task requires each human being to “reconstitute his agency” from the parts of his soul, that is, to “make himself into a particular person” (SC, p. 130). Is this value—that first and foremost we must fulfill this task—something we create by imposing universal law on some desire or interest? Perhaps Korsgaard thinks that it is. However, she speaks repeatedly as if this value is imposed upon us by our nature rather than being something that we create in this way (for example, see SC, pp. xi–xii, 19–20, 23–24, 131, 175). In any case, the procedure must be followed because that is the only way to be an agent. Hence, that we must be an agent in the first place is a value that must be justified prior to the justification of the procedure. Since the need to follow the procedure is a consequence of the need to be an agent, surely the need to be an agent would not be justified even if it were warranted by the procedure (that would be a very tight circle). Moreover, if acting as such is not sufficient to warrant the normativity of the need to be an agent (in line with the objection above), then it seems that the source of this normativity can only be nature itself. We are left with the conclusion that, at the heart of Korsgaard’s enterprise, there is an unacknowledged—and indeed disavowed—assumption that human nature imposes a normative requirement on us, not because there is any meaningful sense in which we have created, endorsed, or tacitly committed ourselves to it, but simply because we find ourselves confronted with this normative fact. In the terminology of SN, it is not evident that Korsgaard has found a way to defend moral realism that does not depend in the end on a substantive, in contrast to a procedural, realism (for a more recent discussion of this distinction, see CA, chap. 10). Moreover, if this is a normative fact imposed by our nature, then there might be a basis for saying that we ought to respect, and not merely begrudgingly go along with, our humanity. For, on this view, humanity is the source of the most fundamental requirement that we are to respect. But any such contention would face skeptical challenges that she claims it is an advantage of her position to avoid.

That our task is to constitute ourselves as agents is another source of perplexity in Korsgaard’s account. According to this account, at the outset (and, in a sense, at each moment of choice), we are, as it were, potential-agents that consist of not-yet-unified parts and our job is to unify these parts into a whole so that genuine action is possible. But if at the initial stage I am a mere potential-agent, a heap of parts, then what can it mean to say that I turn this potential-agent into an actual agent or that I have the responsibility to do so? Since at this stage there are only parts, it seems that there is not yet an I, an agent, who could perform this task or be held accountable for performing it. It is hard to see how a self, in Korsgaard’s sense, could constitute itself from parts or have the job of doing so. Hence, the Constitutional Model appears to be no more intelligible than the Combat Model.

This difficulty, of course, points to a central dis-analogy in Plato’s analogy of the city and the soul: individual persons can constitute themselves as a collective agent and thereby
form a city because the persons are already agents, but parts of the soul cannot similarly constitute themselves as an individual agent because the parts of the soul are not already agents. For Korsgaard, the issue is most evident in her understanding of deliberation. She says that in deliberation “it is as if there is something over and above all your incentives, something which is you, and which chooses which incentive to act on” (*SC*, p. 72; cf. pp. 75, 126, 134–35). This is a familiar way of thinking about deliberation. However, it seems that for Korsgaard this *you* which chooses must be either (1) one of the parts that constitutes the self or (2) the self that is constituted by the unification of the parts. If (1), then one of the parts is already an agent since it makes choices, and presumably this agent does not constitute itself (if it did, we would be faced with a regress of agents being constituted by further agents). Korsgaard does sometimes speak as if the parts were themselves persons (see *SC*, pp. 153–54), but presumably she is speaking metaphorically. If (2), which I take to be Korsgaard’s considered position, then it needs to be explained what the (potential-agent) part is that is unified with incentives so as to constitute the self, and how it is that the self constitutes itself by unifying this part and the incentives. It might seem that the part is what Korsgaard calls (in the argument against particularistic willing) “the principle of choice” among incentives. However, she appears to deny this when she says that this principle is not some force in the self on a par with the incentives. Rather, as we have seen, the self constitutes itself by regarding the principle of choice as “expressive” of, or “representative” of, itself (see *SC*, p. 75). Only in this way could the self regard itself as the cause of the movements that constitute its actions. But then it would seem that the only parts are the incentives and that the self that constitutes itself is always already, prior to each moment of self-constitution, something capable of such acts of identification. But what could this something be if it is not a self that exists prior to these acts?

Korsgaard acknowledges that self-constitution looks paradoxical (see *SC*, pp. 20, 35), but her resolution of the paradox is not convincing. In her primary response, she argues that a giraffe is constantly making itself into a giraffe and that we see nothing paradoxical in this. Likewise, she says, a person is constantly making itself into a person, and there is no paradox in this either. It is the essence of living things to constitute themselves in this way (see *SC*, p. 42). However, it is misleading to say that, by eating, a giraffe is making itself into a giraffe: what it is doing is maintaining itself as the giraffe that it already is (Korsgaard herself sometimes puts it this way). There is no paradox because potential-giraffe-parts do not have the job of constituting themselves as a giraffe. Rather, as we might put it, an actual giraffe has the job of continuing its existence by consuming food. If there is an analogy here all that can be said is that an actual person has the job of maintaining himself as the person he already is. This is very different than saying that potential-person parts have the job of constituting themselves into an actual person by unifying themselves through action.

Perhaps self-constitution for Korsgaard means nothing more than the activity of maintaining oneself as the person one already is by acting. This would resolve the paradox, but then it is unhelpful and misleading to speak of putting back together the parts of the soul that self-consciousness has created so as to reconstitute our agency (as she puts it in *SC*, pp. 125–26, 130; cf. p. 213). It might also be said that self-constitution simply means the activity of constituting oneself as a person with one particular practical identity rather than another. This too would not be paradoxical, but saying this does not require speaking of parts that need to be reconstituted as a whole. Even if, as Korsgaard supposes, the only way to be a person at all is to be a person with some particular practical identity or another, it needs to be shown how parts that are not yet a person are constituted as a person with a specific practical identity.
However, it is not obvious that Korsgaard needs the concept of self-constitution to mount
the arguments for her major claims. She might have said, more simply, that we are agents
in the sense that we have the capacity for self-government (that is, for efficacy, autonomy,
and morality) that we may exercise well or badly. When we exercise it well, we are true to
our nature as the agents we are. To the extent that we exercise it badly, we are false to our
nature: though we are still agents, we are not functioning as agents should. It is not clear
that the language of constituting ourselves as agents is really necessary.

A major—arguably the—aim of the Republic, and especially of the analogy of the city
and the soul, is to establish that a just person will have a eudaimôn life (see Rep. 367c–e
and 586e–587a). Korsgaard denies this, at least in extreme cases, and she may embrace a
position closer to Aristotle’s, that virtue typically but not inevitably brings about eudaimonia
(see CA, p. 149). The fully just person “on the rack,” she says, is “not necessarily happy,”
but she is “entirely self-governed” and “completely self-possessed” (SC, p. 180 and CA,
p. 120). Korsgaard often writes as if the only thing that ultimately matters for a person is
being a unified agent (for example, see SC, pp. 7, 25–26, 35, 214). If her arguments for the
implications of agency are sound, striving to be a unified agent matters a great deal, for
it commits us to being governed by principles concerning efficacy, autonomy, and moral-
ity. But even if her arguments are sound, surely there is a good deal that matters in human
life—that is important for a eudaimôn life—besides being a unified agent, and not only with
respect to what might happen to us (fortune). On her own account, there are many ways
of being a unified agent, and hence many ways of living a morally permissible life, and it
might well be thought that some of these contribute more to our happiness, flourishing or
well-being than others.

Korsgaard effectively stakes out an understanding of well-being that straddles the divide
between subjectivist accounts (for discussion of pleasure- and desire-based accounts, see
SC, pp. 54–56, 120–22, 167–68, 208–09) and some objectivist accounts (for a critique of the
“realist” account of Scanlon and Raz, see SC, pp. 122–22). However, Korsgaard’s position
puts further pressure on the extent to which she can plausibly maintain a purely proceduralist
account of value. In her discussion of this topic, she says, unsurprisingly, that “it is our own
choices that ultimately confer value on objects” (SC, p. 123). This is her proceduralism: as
long as a choice involves a commitment to a universal law, the choice is valid. From this
perspective, the only constraint is formal. However, Korsgaard also says that our choices
respond to features of those objects that we take to give them value, not in an absolute sense,
but in relation “to human physiology and psychology.” At the root of our inclinations, she
says “is a basic suitableness-to-us that is a matter of nature and not of reason” (SC, p. 122;
cf. pp. 120, 208–09). Moreover, she embraces what she calls an Aristotelian conception
of happiness according to which “happiness rests in the excellent activity of our healthy
faculties” (CA, p. 99n27).

Korsgaard does not elaborate on these remarks, but they make it clear that she has aban-
doned Kant’s subjective understanding of happiness. In fact, they would seem to suggest
that there are two criteria of valuable human activities: the categorical imperative and being
suitable to human nature (and possibly, in addition, being suitable to one’s own individual
nature). Hence, even if I am willing to endorse as a universal law that I should pursue act A
for the sake of end E, this would not be worth doing if the end was not suitable to human
nature, such as by being contrary to it or irrelevant to it (as in well-known examples such
as Rawls’s grass counter). For the most part, Korsgaard directly repudiates any non-formal
criterion and insists that we create our values through the application of formal criteria. But
the remarks just quoted appear to tacitly acknowledge that formal criteria may not be enough.
Recall that Korsgaard explicitly acknowledges that she needs, but does not have, a formal prudential principle that tells us how to balance our ends. Hence, both the determination and the balancing of our ends on a formal basis remains something of a mystery in Korsgaard’s account. Even if being a unified agent imposes significant formal constraints on what could count as living well, it is hard to see that these constraints are sufficient to determine the nature of well-being. Korsgaard is enough of an Aristotelian to suspect this, but she is too much of a Kantian to acknowledge the limitations this imposes on her formalist approach. A theory of well-being is thus a major lacuna in her constitutivist position. This is a significant omission in view of her ambition to establish affinities between the autonomy-based moral philosophy of Kant and the eudaimonistic moral philosophies of Plato and Aristotle.¹

¹I have received helpful comments on an earlier draft of this review from John Davenport, Ryan Kemp, and Scott O’Leary.