Practical Reason and the Unity of Agency

Critical Notice

CHRISTINE M. KORSGAARD, *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press 2009. Pp xiv + 230.

Self-Constitution is a thrillingly ambitious book. Ranging widely over both historical and contemporary debates in ethics and the philosophy of agency, Korsgaard sets herself the task of answering some of the hardest questions moral philosophy has to ask. It is required reading for anyone with interests in agency, practical reason, personal identity, or the ethical teachings of Plato, Aristotle, Hume, or Kant. It is also that rarest of academic books: a major contribution to ongoing debate written with such warmth, wit and clarity as to make it accessible to almost any reader. In short, and notwithstanding its failure to fully convince at least this reader of its conclusions, *Self-Constitution* is a splendid piece of philosophy.

The book has three main themes: 'the nature of action, the constitution of personal or practical identity, and the normativity of the principles of practical reason' (1.1.6),[1] though it is the last of these that gives the book its unifying structure. As in her earlier *Sources of Normativity*,[2] Korsgaard sets herself the problem of showing how the principles of practical reason (thought of as including the foundational principles of morality) attain their authority over us. Her answer, which she argues is also that given by Plato and Kant, is that the principles of practical reason are constitutive of agency itself, so that anyone who is trying to act—anyone

[1] All references of this form are to sections of *Self-Constitution*.

[2] C. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).

who is trying to be an agent—is thereby necessarily trying to conform to them. Thus agents are subject to the principles of practical reason simply through their ongoing commitment to being agents. Moreover, we have no choice but to act, for the simple reason that whatever results from choice is an action: 'choosing not to act makes not acting a kind of action' (1.1.1). So we have no choice but to be agents, and therefore no choice but to be practically rational.

The apparent strength of this conclusion calls to mind Robert Nozick's observation about the typical impotence of philosophical arguments:

> Perhaps philosophers need arguments so powerful they set up reverberations in the brain: if the person refuses to accept the conclusion, he *dies*. How's that for a powerful argument? Yet, as with other physical threats ('your money or your life'), he can choose defiance. A 'perfect' philosophical argument would leave no choice.[3]

Korsgaard may be read as trying to provide precisely such a 'perfect' argument. On her view, action is self-constitution, and one constitutes oneself as an agent by choosing in accordance with the principles of practical rationality. Her idea is not that the irrational or immoral agent is somehow forced to mend his ways, say by some sanction of non-existence; the picture is not one in which there *first* exists an agent who *then* chooses irrationality and *as a result* goes up in a puff of smoke. Rather, the entity that 'chooses' irrationality simply fails to constitute itself as an agent in the first place—and so does not choose anything at all (1.4.2). If successful, Korsgaard's argument shows that we quite literally have no choice but to (attempt to) act in accordance with the principles of practical reason. That conclusion is enticingly strong.

---

[3] R. Nozick, *Philosophical Explanations* (Cambridge, M. A.: Harvard University Press, 1981), p.4.

To get to it, Korsgaard faces three major tasks: (1) establishing that the principles of practical reason are constitutive of agency; (2) establishing that the principles of practical reason entail the principles of morality; and (3) explaining how, if this is true, irrational and immoral action are possible. For various presentational reasons Korsgaard tackles (3) in advance of (2), and this discussion follows her ordering. In it I highlight some key moves in her argument that failed to convince, and suggest places where I feel more could usefully be said. My focus throughout is on Korsgaard's problematic—to my mind—notion of agential unity. I begin with (1).

1. The Principles of Practical Reason

According to Korsgaard, to act is to constitute yourself as the cause of some end (4.3.4). Korsgaard draws out two aspects of this seemingly simple analysis, which she takes to correspond to the hypothetical and categorical imperatives. First, action requires constituting yourself as the *cause* of your end, and so requires that you take some appropriate means to it. Second, action requires constituting *yourself* as the cause of your end, and so requires that your behaviour be the result of your own activity and not that of mere forces at work within you. Korsgaard labels these the requirements of *efficacy* and *autonomy*, and takes them to be constitutive standards of agents derivable from the simple analysis above (5.1.1). Here I discuss only her argument for the categorical imperative, or the requirement of autonomy.

As in *Sources of Normativity*, Korsgaard distinguishes the categorical imperative from the moral law: whereas the categorical imperative is 'the law of acting only on maxims that you can will to be universal laws', the moral law is 'the law of acting only on maxims that all rational

beings could act on together in a workable cooperative system' (4.5.5). At this stage of her argument Korsgaard is concerned to establish only the former.

When you will a maxim, you affirm through your action your view that your act is worth doing for the sake of your end, given relevant present circumstances. When you will a maxim universally, you commit yourself to regarding your act as worth doing for the sake of your end in all relevantly similar circumstances. On one plausible view of what it is for something to be a reason, this means that you treat your end as a *reason* for performing your act. That is, the imperative of willing only those maxims that you can will as universal laws is simply the imperative of acting only on the basis of what you take yourself to have reason to do. So, in different language, Korsgaard's task is that of showing that acting on reasons is constitutive of action: that one truly qualifies as an agent only insofar as one acts on reasons. Though many contemporary philosophers of action accept (something like) this claim, few offer explicit arguments for it. This lends particular significance to Korsgaard's argument in favour of it, which she dubs 'the argument against particularistic willing'.

The argument is founded on the idea that human beings are uniquely possessed of a certain type of reflective consciousness, and that this makes human action fundamentally distinctive. For whereas other animals are simply caused to act by their instincts, we have the ability to step back from our desires and to make choices about which ones we wish to be led by. This brings with it a need for some basis on which to make these choices, some principle about how we ought to go about choosing. When human beings act, therefore, they express their commitment to some *principle of choice*. And for Korsgaard, acting on such a principle of choice just is acting on a universal law, at least in the minimal sense that is currently at issue.

More precisely, the argument, as presented in 4.3.3, seems to be that (1) any being that is identified exclusively with its incentives, as opposed to a 'principle of choice', is not an agent but instead just 'a series, a *mere heap*, of unrelated impulses'; (2) the particularistic willer (one who wills maxims non-universally) is necessarily identified exclusively with its incentives; therefore (3) the particularistic willer is not an agent.[4]

Evidently, a distinction between being identified with one's desires or 'incentives' and being identified with a 'principle of choice' is central to this argument. It would appear that Korsgaard wishes to keep principles of choice distinct from incentives, even, presumably, from general higher-order incentives (such as desires that one act on desires of some general type). Yet if this is so, then Korsgaard's sub-argument for (1)—the claim that an agent must identify itself with a principle of choice—does not appear to work. Here is the argument for (1):

> ...suppose you experience a conflict of desire: you have a desire to do both A and B, and they are incompatible. You have some principle that favours A over B, so you exercise this principle, and you choose to do A. In this kind of case, you do not regard yourself as a mere passive spectator to the battle between A and B. You regard the choice as yours, as the product of your own activity, because you regard the principle of choice as expressive, or representative, of yourself—of your own causality. You must do so, for the only alternative to identifying with the principle of choice is regarding the principle of choice as some third thing in you, another force on a par with the incentive to do A and the incentive to do B, which happened to throw in its weight in favour of A, in a battle at which you were, after all, a mere passive spectator. But then you cannot regard yourself

[4] A near identical statement of the argument appears in her paper 'Self-constitution in the ethics of Plato and Kant', in *The Constitution of Agency* (Oxford: Oxford University Press, 2008), pp. 123-4. Both are developments of a line of thought that first appeared in *The Sources of Normativity*, pp. 225-33.

> as the cause of the movements which constitute your action. Self-determination, then, requires identification with the principle of choice on which you act. (4.4.3)

If a 'principle of choice' must be something different from a mere higher-order incentive, then this passage does not succeed in showing that one must identify with a principle of choice. One who identifies with a higher-order incentive experiences it as 'expressive, or representative' of himself, and not merely as 'some third thing' in him—this is, after all, just what it *means* to say that he feels identified with it. Accordingly, he too may regard himself 'as the cause of the movements which constitute [his] action'. Against this, Korsgaard might insist that one cannot identify with what is merely a higher-order incentive on the grounds that it, being just another incentive, is in the end just another force at work within one. This is consistent with her psychological picture on which incentives, belonging to the appetitive part of the soul, always begin life 'out there', whereas principles, belonging to the rational part of the soul, are always 'in here'. Yet at this stage of her argument this is no more than a rhetorical stance; what we need is a reason for thinking that one cannot identify with an incentive.

Korsgaard's idea may be that any agent who feels identified with an incentive (even a higher-order one) is necessarily an agent who has cut short her reflection prematurely and so has failed to achieve sufficient reflective distance from her own impulses. For if such an agent were to continue her process of reflection she would surely notice that this higher-order preference is precisely just 'some third thing', some additional incentive to which she is subject, and then find herself faced with the question whether she should endorse *it*. However, the fact that an agent *could* 'step back' from such a higher-order desire and come to see it as 'external' simply illustrates a general fact about the nature of reflective consciousness, namely that no matter what

inner vantage point one occupies, it is *always* possible to step up to an even higher one from which one can observe the point just vacated. And this fact can be used to undermine Korsgaard's position just as much as her opponents'. For an agent who identifies with a particular principle of choice may step back and come to see *her commitment to this principle* simply as another force at work within her; in deciding what to do, she will then have to identify with some *higher-order principle of choice*, which she may in turn step back from; and so on. Finding oneself trapped in an unending process of hierarchical reflection is something of an occupational hazard of being a reflective agent, and it is left unexplained how identification with choice principles are supposed to possess special immunity from additional identification-undermining reflection.

Korsgaard raises this exact worry in the introduction to *The Constitution of Agency*, noting that a regress threatens since 'it appears that we need a reason to conform to one proposed principle rather than another, and, if that is so, there must be a further principle behind every principle, to give us a reason for conforming to it.' Her response is that 'there need be no such regress if there are principles that are *constitutive* of the very rational activities that we are trying to perform when we take control of our beliefs and of our actions'.[5] This may be right, but it is of no use in the argument against particularistic willing, since the aim of the argument is precisely to demonstrate that conformity to rational principle is a constitutive standard of action. Moreover the truth is surely that, in practice, most of us manage to avoid Hamlet-like fates simply through a willingness to break off reflection and act as best we can whenever the contingencies of the world become sufficiently severe. When we do so, we act on the basis of

[5] *Constitution of Agency*, p. 5.

some volitional element with which we identify and from which we have *not yet* acquired reflective distance. Yet if this is how it works then there is no relevant distinction to be drawn between principles and incentives.

So suppose instead we read (1) merely as the claim that an agent must identify with some higher-order volitional element that has general content. The question now is why we should think this, why an agent cannot identify with a higher-order incentive with particular content, such as a desire to act on *this* desire. Harry Frankfurt gives the example of a mother in conflict about whether to give her child up for adoption; though she concludes that the reasons weigh in favour of doing so, when it comes time to sign the papers she instead throws her weight behind her desire to keep her child—'not because she has reconsidered the matter and changed her mind but because she simply cannot bring herself to give her child away'.[6] As described by Frankfurt, this decision is deeply expressive of her volitional character but is not grounded in reason. She merely acts on her inclination to keep her child, an inclination supported in reflection by an inclination to act on that very inclination. Why must we deny this woman agency?

Korsgaard raises and seeks to address what may be a version of this challenge in a footnote:

> Why can't the particularistic willer keep himself separate from his incentives by saying of each of them in turn, 'I am the one who acts on *that* incentive', as it were, mentally pointing, since he cannot regard the incentive as a type and therefore cannot give it a name? That mental pointing is the problem. For what can he mean by 'that incentive'? Simply, 'the one I am acting on now'. So

[6] H. G. Frankfurt, 'On the Necessity of Ideals', in his *Necessity, Volition and Love* (Cambridge: Cambridge University Press, 1999), p. 111.

> his thought would be 'I am the one who is acting on the incentive I am acting on now'. Obviously, the thought lacks content. (4.4.3)

But this leaves it unclear why the particularistic willer cannot regard her incentive as an instance of a type. The idea may be that such a willer would then be committed to desiring the type, and so something general after all (her desire being thus the result of a little syllogism: I want to act on desires of type X; here is a desire of type X; *so* I want to act on this desire.) But surely one may conceive something as an instance of a general type whilst desiring just that instance. Thus the mother Frankfurt describes wishes simply that *she* be moved *now* by her desire to keep *this* child, without thereby wishing that other mothers be moved by desires of that type, or that she be moved by a similar desire if in some parallel situation (perhaps concerning a future child). Theoretically this is no more troublesome than, say, her child's desire for *his blanket*, which may be thought under the general concept of 'blanket' but is nonetheless *desired* only in its full particularity, such that no wish for any other (even exactly similar) blanket forms any part of it. Thus an exclusively particularistic *willer* need not also be an exclusively particularistic *thinker*.

Part of the trouble here is the tantalising brevity with which Korsgaard presents her argument against particularistic willing. It seems that she must say more if she is to explain convincingly why a reflective agent cannot be identified with her incentives, and so why a particularistic willer cannot constitute herself as an agent.

2. The Problem of Bad Action

Suppose that Korsgaard has succeeded in showing that the principles of practical reason are constitutive standards of agency. How does this explain their normativity? Well, anyone

who is trying to act in any way is, trivially, trying to constitute herself as an agent (i.e., as a thing that acts), and an agent *just is* something that operates in accordance with the principles of practical reason. So anyone who is trying to act is necessarily trying to act in accordance with the principles of practical reason. And insofar as such a person acts irrationally, she performs a defective action, and she fails as an agent. To employ Korsgaard's helpful analogy, compare: a house *just is* something that meets certain constitutive standards (for instance, it has walls and a roof, it provides shelter against different types of weather, and so on). So anyone who is trying to build a house is necessarily trying to build something that meets these standards. And insofar as such a person builds something that fails to meet these standards, she fails as a housebuilder. The constitutive standards of houses are normative for anyone trying to make a house, and the constitutive standards of agents are normative for anyone trying to make an agent—which is, trivially, what we try to make of ourselves every time we try to act (2.1).

This line of thought is subject to a well-known objection, which is that it makes a mystery of irrational action. Irrational action is action in violation of the principles of practical reason. But if these principles are constitutive of action, then nothing that is an action can be in violation of them. So, if the principles of practical reason are constitutive of action, then irrational action is impossible. But irrational action is clearly possible. So the principles of practical reason cannot be constitutive of action.

Korsgaard's housebuilding analogy illuminates how this objection might be met. Suppose someone argued as follows: 'A bad house is one that is in violation of the principles of house quality. But if these principles are constitutive of houses, then nothing that is a house can be in violation of them. So bad houses are impossible.' Clearly something has gone wrong in

the argument, since houses are subject to constitutive standards, yet bad houses are possible. What has gone wrong is a failure to see that house quality comes in degrees, and that a defective house becomes no house at all only at the limit. There are many intermediate stages in which a house may be defective but nevertheless still a house. This is the line Korsgaard takes with regard to action. Only at the limit, when behaviour entirely fails to conform with the principles of practical reason, do we deny that it is action. This leaves a large range of *somewhat* irrational action that comes close enough to meeting the standards of practical reason for us to count it as action while nevertheless counting as defective with respect to those same standards.

For this to fly, however, there must be a clear sense in which conformity to the principles of practical reason comes in degrees. Yet it is difficult to see how this is possible. Take the categorical imperative, with which Korsgaard seems most concerned. The problem is that, in light of her argument against particularistic willing, it would seem that one must will one's maxims as universal laws *on pain of being a particularistic willer* (and hence just a 'mere heap' of impulses); that one must identify either with some principle of choice or else with one's particular incentives, with no obvious middle ground. Thus the framework introduced to help facilitate the derivation of the categorical imperative from the concept of agency may not be especially suited to revealing any kind of sliding scale as regards our conformity or nonconformity with that principle.

It is in part to address this problem that Korsgaard introduces the idea of agential unity. Her idea is that some principles of choice unify our agency better than others, and that part of our task as agents is to unify ourselves. This, the standard of agential unity, is indeed something that we may approach more or less closely. Yet the idea of agential unity is introduced in a way

liable to confuse the unwary reader. Previously, the autonomous agent was understood to be the agent who chooses his maxims in accordance with some principle of choice, who acts in accordance with some (that is, *any*) self-chosen practical law. This is what the argument against particularistic willing is supposed to teach us. But now we are presented with a new, stronger conception of the autonomous agent as one who chooses maxims in accordance with a principle of choice that is successful in unifying his agency. Accordingly, we require some stronger parallel of the argument against particularistic willing that will show that agential unification (in this sense) is *also* a constitutive standard of agency. Yet any such argument is presented only implicitly.

Korsgaard first introduces the idea of agential unity as part of an elaboration of the idea that one who follows the categorical imperative constitutes himself as an agent in a way that the particularistic willer does not. The idea is familiarly Platonic and goes as follows. We all agree that there is a difference between a member of a group acting and the group itself acting. So, for instance, if a Canadian political pundit demands that Iran stop developing its nuclear technology, that is just a private individual mouthing off; but if the Minister of Foreign Affairs demands that Iran stop developing its nuclear technology, that is a demand being issued by Canada. And what authorises the Minister of Foreign Affairs to speak on behalf of Canada as a whole is the fact that Canada is constituted by a particular set of rules (its constitution) that set out the roles and procedures necessary to make group agency possible, and the Minister of Foreign Affairs occupies a role that permits him to speak about foreign affairs on behalf of the whole. Analogously, says Korsgaard with Plato, there is a difference between a part of a person acting and the person herself acting. So if a desire bypasses the agent's rational faculty and causes

action directly, that is just a force acting within her; but if she rationally reflects on the desire and wills it, the resulting action is one performed by the agent as a whole. And, importantly, this is not because she identifies with her rational faculty and her rational faculty has just won a battle against desire. It is because she has a sort of internal constitution that alots determinate roles to each of her parts, and it is reason's job to give the agent's final assent to deliberative proposals.

For Korsgaard, this 'shows us why certain formal principles—the categorical imperative, and Plato's principle of justice—are constitutive principles of action: because they bring the constitutional unity that makes action possible to the soul' (7.5.4). As we may recall, for Korsgaard the 'categorical imperative' is simply the imperative of acting in accordance with some principle of choice as opposed to identifying oneself wholly with every passing inclination. One who identifies with a principle of choice is one who has a constitution and is therefore capable of acting as a unified whole. The particularistic willer, by contrast, is like an anarchic group: all sorts of things may be caused by its parts, but there is nothing that counts as an action of the whole. This is why the particularistic willer does not count as an agent.

This is all well and good, but it does not yet offer us any progress regarding the problem of bad action.[7] To explain the possibility of bad action, Korsgaard needs to show how conformity with the categorical imperative can come in degrees. Yet the city/soul analogy does little to help with this: after all, it is not obviously unreasonable to think that, for any group of people, either they are bound by a shared constitution or they are not. At the very least, Korsgaard would need to convince us that 'having a constitution' is something that comes in

[7] Nor does it, being a mere analogy, constitute any independent argument for Korsgaard's claim that the categorical imperative is a constitutive standard of agency.

degrees by providing an account of constitution-possession, *and then* show how this can be appropriately analogised to the personal case. Yet she does neither.

What she does do, instead, is subtly shift the topic of her discussion from the existence of constitutions to their quality. Now, the quality of a constitution is clearly something that can come in degrees, and Korsgaard argues that the quality of a constitution is a matter of the extent to which it unifies one's agency: 'since the aim of the constitution is to unify the soul, the defective constitutions must lead to disunity and to that extent must undercut agency' (8.2.3). So different constitutions bring with them different degrees of unity. But we now have on the table a new and more substantive sense of 'agential unity' than the procedural one described above. In that first, thinner sense, agential unity was simply a matter of *having* a constitution—*any* constitution—and so having parts that are authorised to 'speak for' the whole. Now we are invited to consider a *further* type of agential unity for which this bare procedural unity could not suffice.

Korsgaard may see this more substantive unity as occupying a point on the same scale as the thin procedural unity (with thin unity merely occupying the scale's lowest point). I am not convinced that the difference between these notions of unity is simply one of degree. However, this disagreement is of little importance, for either way there are two things that it is incumbent on Korsgaard to demonstrate. First, she must show that there is a genuine sense in which different principles of choice work to unify one's agency to different extents, and in which conformity to the moral law unifies one's agency the most. Second, she must show that agential unity *in this same sense* is a constitutive standard of action. This requires a fresh argument, since the argument against particularistic willing sought to establish only that agential unity *in the*

minimal sense of conformity with the categorical imperative is a constitutive standard of action. I will now say a little more about this second task before focusing on the first in the final two sections.

Does Korsgaard really need to spell out a new argument to show that agential unity is a constitutive standard of action? After all, a disunified house—a pile of bricks and cement—is not a house, and wanting to create a house is necessarily wanting to create a unified house. By analogy, a disunified agent—a bundle of incentives—is not an agent, and wanting to make oneself into an agent is necessarily wanting to make oneself into a unified agent. Yet matters are not so straightforward. Just because houses have unity as a constitutive standard does not mean that agents do. This is not to deny that agential unity is a good thing, or that we have ample reasons to prefer unity to disunity as regards our own agency. This may be true without unity being a *constitutive* standard of agency. It might be, simply, that there are unified and disunified agents, just as there are imaginative and unimaginative agents, clever and stupid agents, agile and clumsy agents, and so on. That is, unity may be merely an *external* standard of agents.

At one point (2.1.1) Korsgaard suggests that she thinks *every* object has unity as a constitutive standard, and perhaps there is a sense in which this is true. Yet note that there are different types of unity, so that one and the same thing may be unified in one respect and disunified in another. For instance, a terrorist organisation may be ideologically unified but structurally and geographically disunified. So while it may be true that everything must have unity *in some sense* as a constitutive standard, it is certainly false that everything must have unity *in every sense* as a constitutive standard. And that means that it is an open question whether

unity in the more sophisticated sense that Korsgaard now introduces is a constitutive standard of agency.

To see this clearly, consider the following two obviously fallacious arguments:

(1) A deconstructionist house, like Frank Gehry's Santa Monica home, is expressly designed to manifest a type of disunity amongst its parts. Yet unity is a constitutive standard of houses. Therefore Gehry lives in a defective house.

(2) Republican constitutions are expressly designed to maintain separations of powers, that is, a type of disunity amongst the institutions that they govern. Yet unity is a constitutive standard of political constitutions. So republican constitutions are defective.

Both arguments involve equivocations on 'unity': though unity is in some sense a constitutive standard of both houses and constitutions, it is not such a standard in the senses employed in the arguments' first premises. Accordingly, it is not enough for Korsgaard simply to show that unity is in some sense a constitutive standard of agency, and that the moral law in some sense guarantees the unity of one's agency. She must also show that this is the same sense.

3. Agential Unification

So what precisely does Korsgaard mean by 'agential unification'? Perhaps a good place to start is with Korsgaard's ideas concerning what an agent *is*. As we have seen, for Korsgaard the agent is not identified with any particular set of attitudes, but rather with its commitment to some particular principle of choice. The idea is that the agent is something that stands over and

above its incentives, in much the same way as the constitution of a city stands over and above its population (7.1.3). If this is what the agent is, however, it may be hard to see what its unification could consist in—after all, it is doubtful whether a mere commitment is even something that has parts. Yet it must be recalled that, for Korsgaard, action is self-constitution, and so anything that counts as an agent has *already* been unified. The work of self-constitution is that of taking the various parts of the soul (reason, appetite, and so on) that have been divided by our reflective consciousness and of reintegrating them into a unified whole. So what stands in need of unification is not the agent itself, for that is already the finished product. What stands in need of unification is the (as it were) *proto*-agent, the 'mere heap' of incentives.

However, as we have seen, there is a sense in which this proto-agent can be unified by a commitment to *any* principle of choice, by the adoption of *any* constitutional framework that allows for the type of procedural unity required for one of the parts to count as speaking for the whole. We need then a more precise specification of the more substantive unity to which Korsgaard must also appeal in order to make work her solution to the problem of bad action. So perhaps we will better understand what Korsgaard has in mind here by looking to her examples of substantive agential *disunity*. The most developed of these is that of Jeremy, the democratic soul:

> Jeremy, a college student, settles down at his desk one evening to study for an examination. Finding himself a little too restless to concentrate, he decides to take a walk in the fresh air first. His walk takes him past a nearby bookstore, where the sight of an enticing title draws him in to look at the book. Before he finds it, however, he meets his friend Neil, who invites him to join some of the other kids at the bar next door for a beer. Jeremy decides to have just one, and he goes with Neil to the bar. While waiting for his beer, however, he finds that the loud noise in the bar gives him a

> headache, and he decides to return home without having the beer. He is now, however, in too much pain to study. So Jeremy doesn't study for his examination, hardly gets a walk, doesn't look at the book, and doesn't drink his beer. (8.3.4)

Jeremy has had a frustrating evening. But in what sense is he *disunified*? According to Korsgaard, the problem is that 'each of Jeremy's impulses leads him to an action that completely undercuts the satisfaction of the last one', rendering him 'almost completely *incapable of effective action*' (8.3.4). His problem then is with *efficacy*, which suggests that Jeremy is failing properly to follow the hypothetical imperative. For Korsgaard, willing an end involves committing oneself to taking the necessary means to that end, unless one finds reason for revising it; the hypothetical imperative commands us to live up to these commitments (4.3.4).[8] The problem, however, is that Jeremy *does* live up to these commitments. Being a democratic soul, his principle of choice is to act always on his strongest desire, that is, to treat all of his desires as reasons.[9] So he sets himself the end of studying, but when he more strongly desires fresh air he takes this as a conclusive reason for *abandoning* his original end of studying. In this way his walk represents no violation of the hypothetical imperative, since he no longer has the end that it impedes.

Jeremy, we may presume, is wholly committed to his higher-order principle of always acting on whatever happens to be his strongest current inclination. And being committed, as he is, to the categorical imperative, we may presume that he wills only those maxims that he can

---

[8] See also pp. 56-60 of her 'The normativity of instrumental reason', in *The Constitution of Agency*.

[9] That is to say, the democratic soul is *not* a particularistic willer. 'Someone who takes "I shall do the things I am inclined to do, simply because I am inlined to do them" as his maxim has adopted a universal principle, not a particular one: he has the principle of treating his inclinations *as such* as reasons' (4.4.3).

commit himself to acting on in the future. That is, what he wills as a universal law is not the maxim 'I will study this evening in order to do well in my exam', but the maxim 'I will study this evening in order to satisfy my current strongest inclination, which is to do well in my exam'. And his commitment to *this* maxim is not undermined when his current strongest inclination changes. So Jeremy suffers from no deficiency of commitment to his maxims (indeed, this is precisely what distinguishes him from the particularistic willer). What he does suffer from is a lack of commitment to his various more substantive ends: passing his exam, acquiring a book, and socialising. This is because his only real commitment is to doing whatever he wants to do at the time.

Yet this lack of more substantive commitment cannot be where his problem lies either, since the only soul that is free of *this* deficiency of commitment, who is committed come what may to a substantive end, is the one soul supposed to be even *less* unified than Jeremy: the tyrant. The tyrant is one who pursues power, say, with ruthless single-mindedness, letting nothing throw him off course. By contrast, the aristocratic soul is presented as one who is always willing to reconsider her standing commitments in light of relevant new circumstances. (This is what Korsgaard means when she says that the maxims of the autonomous agent are 'provisionally universal'.[10]) So if she makes you a promise, and then discovers that her breaking that promise is necessary for the survival of the entire world, we are not simply to assume that she will remain

[10] '...there's no general reason to suppose we can think of everything in advance. When we adopt a maxim as a universal law, we know that there might be cases, cases we haven't thought of, which would show us that it is not universal after all. In that sense we can allow for exceptions. But so long as the commitment to revise in the face of exceptions is in place, the maxim is not merely general. It is provisionally universal.' (4.4.2)

stubbornly committed to keeping her promise. Of course, the aristocrat *is* wholly committed at the higher-order level: she is stubbornly committed to doing what is just, that is, to her principle of choice. But so is the democrat, and in neither case does the strength of this commitment entail any more substantive commitment lower down.

To draw this out, note that even an aristocrat may end up in Jeremy's situation. Suppose that the book is a rare edition that Jeremy has been trying to find to give as a present to his father, and that Neil is an important love interest. Thus Jeremy has three standing projects—doing well in his exam, bringing joy to his father, and winning Neil's affections—which on this frustrating night are rendered unexpectedly incompatible by an uncooperative world. Finding it hard to study, he decides to clear his head so as to better pursue his project of doing well in his exam. Seeing the book in the window brings two of his projects into conflict, and on reflection he decides to delay his studies in order to take advantage of the rare opportunity. Being invited for drinks by Neil then brings all three of his projects into conflict, and on reflection he decides to sacrifice his studies and risk losing the book in order not to rebuff Neil, before ending up having to do this anyway when he becomes unwell. I find nothing in Korsgaard's conception of the just person that guarantees immunity from this sort of fiasco. Sometimes even the virtuous fail to get things done.

Korsgaard writes:

> On certain occasions, the people with the other constitutions fall apart. For the truly just person, the aristocratic soul, there are no such occasions. Anything could happen to her, anything at all, and she will still follow her own principles—and that is because she has universal principles, principles that can consistently be followed in any kind of case. (9.1.5)

Yet we are still no closer to understanding in what sense those with non-aristocratic constitutions 'fall apart'. They also have universal principles (insofar as they obey the categorical imperative, which is both necessary and sufficient for having *any* constitution), and they can consistently follow them in any kind of case: no matter what happens, the timocrat will follow his principle of doing what is honourable, the oligarch that of doing what is prudent, the democrat that of doing whatever he wants, and so on. (This of course does not mean that they will all *succeed* in doing what is honourable, in satisfying their desires, and so on—the world can be a frustrating place. But nor is there any guarantee that the aristocrat will always succeed in doing what is just (9.1.3).)

4. The Moral Law

In the final two chapters of the book, Korsgaard turns to the task of demonstrating that not only the categorical imperative but also the moral law is a constitutive standard of agency. To get from the categorical imperative to the moral law, Korsgaard must show that (1) 'the domain over which the universal law ranges must be rational beings as such', and (2) 'the reasons embodied in universal maxims must be understood as public' (4.5.5). Instead of trying to do justice to her complex arguments for these two claims, in this last section I look briefly just at one line of thought that might shed further light on the idea of agential unity that we have been considering.

Korsgaard explains clearly why (1) is insufficient by itself to commit us to the moral law:

> Suppose you and I are competing for some object we both want. I think I have a reason to shoot you, so that I can get the object. On the private conception of reasons, universalisability commits me to thinking you also have a reason to shoot me, so that you can get the object. I simply acknowledge that fact, and conclude that the two of us are at war. Since I think you really do have a reason to shoot me, I think I'd better try very hard to shoot you first. (9.4.5)

So to establish the moral law we must also show (2), that agents are required to treat their reasons as public. For Korsgaard, 'public reasons' are reasons 'whose normative force can extend across the boundaries between people', and are 'roughly the same as what are sometimes called objective, or agent-neutral reasons' (9.4.5).

Korsgaard argues for (2) on the grounds that treating one's reasons as public is necessary for maintaining one's diachronic unity: an agent will be diachronically disunified unless she regards the reasons of her past and future selves as having normative force for her in the present, and hence as more than merely private reasons; in this way, 'shared normative force is the glue that holds an agent together' (9.7.4). The central example is that of Derek Parfit's Russian nobleman, who plans as a young socialist that he will distribute his inheritance to the peasants once he receives it, but also predicts that he will grow more conservative over time and worries that he might then decide to keep it for himself. So he contracts now to give away his estates when he gets them, a contract that only his wife can revoke, and he makes his wife promise not to revoke it even if he asks her to in the future.[11] Korsgaard diagnoses the nobleman's problem as being that he treats his reasons as private and so, like the parties to the stand-off described above, gets himself into a state of essential conflict with his future self. She writes:

---

[11] D. Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1986), pp. 327-8.

> He doesn't think of his future reasons as reasons—he thinks of them as facts to contend with, as tools and obstacles, and in his case mainly obstacles—and he is therefore in a condition of war with himself. His efforts as a young man are dedicated to ensuring that his younger self wins, and his older self loses. His soul is therefore characterised by civil war, and that is why he fails as an agent (9.4.10)

Korsgaard seems to understand the young nobleman as having (what he regards as) a private reason to ensure that the estates end up with the peasants, and as predicting that his older self will have a private reason to keep the estates. Being private, this future reason has no normative force for him, and so he simply concludes that they are in conflict and sets about trying to win. This is the source of his disunity.

But what is the ideal of unity with which the nobleman is to be contrasted? Clearly, a person who did not suffer from such a radical shift in fundamental values would count, in one straightforward sense, as more unified. Equally clearly, this cannot be what Korsgaard has in mind, since present commitments to public reasons, universalisability, and the moral law provide no guarantee against such contingencies. Even a supremely virtuous agent may come to believe, with good evidence, that she will later suffer some catastrophic moral decline, losing her ability to respond to some central class of moral reasons. An agent in such a predicament may then have a duty to act now in ways that will minimise the damage she expects her future self to cause, thus entering into diachronic conflict with herself. This is a clear form of diachronic disunity, but it is an empirical disunity based on a brute discontinuity of values and capacities, and so not an illness curable by any Korsgaardian prescription.

So even Korsgaard's ideally unified agent might come to have good reason to plot against her future self. What is supposed to distinguish her from Parfit's nobleman, it seems, is that she is committed to regarding such future changes as instances of *rational decline*: Parfit's nobleman, we are told, 'does not anticipate that he is going to become irrational', but instead 'simply believes that when he is older he is going to have different values' (9.3.2). Korsgaard suggests that this failure to accord normative standing to the views of his future self is what renders him disunified (9.7.2). If he were to regard his reasons as public, he would be committed to judging the behaviour of his future self against the standard provided by those reasons.

Yet the nobleman does not need to think of his reasons as public in order to be committed to regarding his future conduct as irrational. The right kinds of private reasons can do the job just as well. For instance, consider a variation on Parfit's case in which the nobleman's exposure to socialist ideas has had the effect of merely radicalising his standing notions of *noblesse oblige* and feudal care, so that he now takes himself to have reason to distribute his inheritance to *his* serfs. This reason is private and agent-relative; universalising, it commits him simply to the view that each nobleman has reason to give his inheritance to his (i.e. that nobleman's) serfs. *A fortiori*, it commits him to the view that his future self has reason to give his inheritance to his serfs. Insofar as he expects his future self to fail in this, he expects his future self to suffer from a failure of rationality. So even if a commitment to diachronic rational self-assessment is

necessary for some type of agential unity (some manner of rational unity, perhaps), it is far from obvious why unity in this sense is uniquely achieved by the aristocratic soul. [12]

Moreover, even were it to be so uniquely achieved, it would still remain to be shown that this sort of rational unity is more than a merely external standard of agents; that it is constitutive of agency, so that only agents unified in this way can constitute themselves as causes of ends. It may seem obvious that a disunified agent cannot constitute *herself* as the cause of an end, since there is no single entity to act as a cause. But, again, there are various senses in which an agent might count as 'disunified', and Korsgaard's claim is not that each and every one of these types of disunity necessarily undoes agency. For instance, the merely empirical diachronic disunity mentioned above, involving brute changes in desires, is not in itself supposed to scupper the virtuous being's claims to agency. So an explanation is still required as to why this sort of unity counts merely as an external standard of agency (if indeed it is any standard at all), whereas Korsgaard's favoured type of unity counts as a constitutive standard.

5. Conclusion

I have voiced concerns regarding two links in Korsgaard's impressive chain of reasoning, a chain intended to tie the normativity of the moral law at one end to the idea of action at the

---

[12] Korsgaard observes that 'a private reason is like a toothbrush. They are all pretty much alike, but we must each have our own' (9.4.5). So when I take myself to be bound by a private reason, universalisability may commit me to taking you to be bound by a similar reason of your own. And in failing to act on that reason you manifest a type of irrationality that I am bound to recognise as such—not because you fail to act on something that has normative force for me, but because you fail to act on something that has normative force for you.

other. The first of these concerned her argument against particularistic willing, an argument designed to derive the authority of the categorical imperative from the very idea of reflective action, by showing that one truly qualifies as an agent only insofar as one acts on a 'principle of choice'. I found it unclear what distinguished a principle of choice from a higher-order incentive with general content, and (if there is no such distinction) what argument was intended to show that the generality of the incentive's content is necessary for action.

The second concern was about Korsgaard's attempted solution of the problem of bad action. Korsgaard's basic thought about how to deal with this problem—that behaviour may be closer or further from the rational ideal and so action merely to some extent—is promising. Yet her elaboration of this thought in terms of more and less agentially unifying practical principles is hindered by her employment of an insufficiently defined notion of agential unity. Too often talk of 'unity' and 'unification' seems to have predominantly rhetorical force; if the necessary conceptual connections—between agency and unification, and between unification and rationality—are to be established, a more precise articulation of the central notion is required.

No single critical note could hope to do justice to the full scale and ambition of Korsgaard's enjoyable book. Indeed, entirely unmentioned are her discussions of the hypothetical imperative, teleology and the idea of function, animal agency, prudence, practical identity and joint agency, together with her various criticisms of the 'empiricist' and 'dogmatic rationalist' approaches to practical reason. Also undiscussed are her often insightful readings of Plato, Aristotle, Hume and Kant. On all of these topics *Self-Constitution* has given me plenty to think with, and I doubt I will be the only reader left in such a happy position.