Self to Self

Selected Essays

J. DAVID VELLEMAN
New York University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521670241

First published 2006
Reprinted 2007

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Velleman, James David.
Self to Self : selected essays / J. David Velleman.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-85429-6 (hardcover) – ISBN 0-521-67024-1 (pbk.)
1. Self. 2. Self (Philosophy) 3. Kant, Immanuel, 1724–1804 – Ethics. I. Title.
BF697.V45 2005
126–dc22 2005008114

ISBN 978-0-521-85429-0 Hardback
ISBN 978-0-521-67024-1 Paperback

Transferred to digital printing 2009

12

Willing the Law

Kant believes that we must come up against practical conflicts in order to feel the normative force of morality, because that force consists in our own unwillingness to live with practical conflicts of two kinds: contradictions in conception and contradictions in the will. Every instance of immorality is, according to Kant, an instance of one or the other conflict; and only by recognizing and recoiling from these conflicts do we come under the guidance of morality. Because these conflicts are contradictions, they are conflicts of reason, and their instances are irrational as well as immoral. We come under moral guidance, then, in recognizing and recoiling from conflicts of practical reason.

I am going to argue against Kant's account of contradictions in the will, and in favor of an alternative account, which I shall call "concessive." My arguments will imply that Kant is wrong about one of the ways in which wrongdoing is irrational, and hence about one of the ways in which we are guided by morality.

This chapter originally appeared in Baumann, Peter, and Betzler, Monika (eds.), *Practical Conflicts: New Philosophical Essays* (Cambridge: Cambridge University Press, 2004), 27–56. It is reprinted by permission of Cambridge University Press. The chapter develops a suggestion that I make at the end of "The Self as Narrator," a paper on Dan Dennett's conception of the self (Chapter 9 in the present volume). Audiences to which I presented that paper have helped me to write this one; they include the philosophy departments at the University of Pittsburgh, the University of Maryland (College Park), and the University of Chicago. This paper was the target of a critique by Jürgen Müller, delivered to the Göttinger Philosophisches Kolloquium in January 2003, where much helpful discussion ensued. I am also grateful to Jerry Cohen, Tamar Schapiro, Nishiten Shah, and Ralph Wedgwood for comments on earlier drafts.

Kant is committed to the proposition (i) that wrongdoing entails irrationality in the agent, since a perfectly rational agent always does the right thing. He is also committed to the more specific proposition (ii) that wrongdoing entails irrationality in the action, since the balance of valid reasons for acting always favors doing the right thing. The latter, more specific proposition has often been the target of criticism.[1] The reasons there are for an agent to act seem to depend on aspects of his circumstances and psychological makeup that cannot be guaranteed to harmonize with what's right. A particular agent can therefore be a "hard case" in the sense that the right act is one that he has no reason to perform.[2] A proposition to which Kant is committed thus appears to be false.

In the debate over this proposition, Kantians have pointed out that a person can indeed be a hard case in the sense that he is not moved by reasons for him to do right; but in that instance, he is not exempt from those reasons but rather irrationally insensitive to them.[3] What depends on the agent's psychological makeup, then, is whether he is rational in responding to reasons for doing right, not whether such reasons apply to him.

Although I have in the past seconded this response to the critics of Kantianism,[4] I am also tempted to make a more concessive response. I am tempted to concede that an agent may do something wrong, not because he is insensitive to reasons for doing right, but because he has no such reasons. Yet having conceded that an agent can lack sufficient reason for doing the right thing, I would insist that such an agent is nevertheless irrational. I am therefore inclined to assert proposition (i) but deny (ii). The resulting view is what I shall call "concessive Kantianism."

My goal in this chapter is not so much to defend concessive Kantianism as to explain it and to show that it may in fact be implicit in a prominent reconstruction of Kantian ethics. I'll begin by explaining how an immoral act can be rational in itself while being the act of an irrational agent. The explanation will be that an agent can be irrational by virtue of having a problematic set of reasons for acting, even though he proceeds to take the course of action favored by the balance of those reasons. The result is a rational act performed by an irrational agent.[5]

[1] See, e.g., Foot 1978a, 1978b; Williams 1981a, 1995.

[2] The phrase "hard case" comes from Williams 1995: 39.

[3] See Korsgaard 1986.

[4] See Velleman 1996.

[5] A similar thesis is defended by Michelle Mason in her doctoral dissertation, "Moral Virtue and Reasons for Action" (2001); see esp. ch. 2.

This explanation commits me to evaluating an agent as rational or irrational on the basis of the reasons that he has for acting. It therefore commits me to holding an agent responsible for the reasons available to him. After offering a rather breezy defense of this commitment, I'll point out that it originates in the moral psychology of Kant's *Groundwork*. Indeed, Kant himself is committed to holding an agent responsible for his reasons in an especially rigorous way, and here is where my version of Kantianism makes its characteristic concession. I'll try to explain why my concessive way of holding an agent responsible for his reasons should be preferred to Kant's. Finally, I'll argue that this concessive version of Kantianism is implicit in the reconstruction of Kantian ethics recently offered by Christine Korsgaard in the symposium on her Tanner Lectures (1996d).

The upshot will be a novel account of contradictions in the will – the second and, I think, less obvious kind of practical conflict that we are enjoined to avoid in Kantian ethics. I call the account novel not to boast but to concede that it is historically inaccurate. Then again, maybe Kantian ethics could do with a little less historical accuracy.

An Irrational Sort of Person

Let me turn, then, to my exposition of the view that I call concessive Kantianism. And let me begin by illustrating the view with an example borrowed from Bernard Williams:[6]

> Suppose, for example, I think someone . . . ought to be nicer to his wife. I say, "You have a reason to be nicer to her." He says, "What reason?" I say, "Because she is your wife." He says – and he is a very hard case – "I don't care. Don't you understand. I really do not care." I try various things on him, and try to involve him in this business; and I find that he really is a hard case: there is *nothing* in his motivational set that gives him a reason to be nicer to his wife as things are.

Here an orthodox Kantian may insist that the man doesn't need his "motivational set" to give him a reason for being nicer: he already has plenty of reasons. What his motivational set must give him is a motive responsive to those reasons, a motive in the absence of which the agent counts as irrational. The possibility that I am now entertaining, however, is that the agent may not have any reason for being nicer to his wife, and

[6] 1995: 39. I have omitted a parenthetical remark that the "ought" in this passage is used "in an unspecific way" – which means, I take it, a way that isn't specifically moral. I have followed Williams in this respect by speaking of actions as "right" and "wrong" in senses that aren't necessarily moral.

this because of his motivational profile. Given the sort of man he is, he may in fact have no reason to be nicer. But the sort of man who has no reason to be nice to his wife, I want to say, is an irrational sort of man to be.

How can someone be irrational when he is nevertheless acting on the balance of reasons that apply to him? How can he be irrational for failing to have the right reasons? The answer is that there is more than one way to be irrational.

On the one hand, a person is irrational if he lacks some capacities or dispositions that are essential to the activity of practical reasoning. If someone lacks the ability to recognize which considerations are the stronger reasons for him to act, or a disposition to be guided by such considerations, then he is deficient as a practical reasoner and hence irrational. On the other hand, a person can be irrational because his situation or personality presents him with reasons that hinder practical reasoning, without necessarily undermining his capacities as a reasoner.

Consider a person who is torn between two conflicting projects. He aspires to great wealth and success, for example, while also seeking a simple life of reflection and self-cultivation. He may be perfectly capable of weighing the reasons that issue from these ideals, and perfectly responsive to the force of those reasons. Indeed, long experience with difficult choices may have made him unusually adept at the art of deliberation. Yet there is something irrational about being so conflicted, about holding on to goals that cannot be jointly attained.

This example, underdescribed though it is, suggests that Williams's hard case has been described even less adequately. Not even the most demanding Kantian would balk at the idea of two people's having no reason to be nice to one another, even if they happen to be married. The Categorical Imperative doesn't require that everyone be nice to everyone else, and marriage is a context in which people can lose their reasons for being nice. But in such cases, people have usually lost their reasons for being married, or for living together in circumstances that provide opportunities for being nice or the reverse. What Williams's example invites us to imagine, I think, is a case in which a man isn't as nice to a woman as he should be in light of their remaining together as husband and wife. For some reason – and we imagine that the man has a reason – he stays in the marriage while treating his wife as would be appropriate only for a stranger or even an enemy. And now we have imagined an agent who is, in some way that remains to be described, committed to conflicting projects. Something in his life gives him reason to be married

to a woman to whom he is unsympathetic or even hostile, and so there must be an underlying practical conflict of some sort.

One might argue that the man is failing to act on the reasons that apply to him, because each of his conflicting projects gives him reason to abandon the other. But reasons for abandoning a project tend to undermine the reasons that issue from it, and so the agent's reasons for abandoning either project are undermined by his reasons for abandoning the other. The man's problem is not that he's inappropriately hostile in light of his commitment to the relationship, since he has reason to give up the relationship in light of his hostility; nor is the problem that he's inappropriately committed to the relationship in light of his hostility, since he has reason to give up the hostility in light of the commitment. His problem is that he has gotten himself into a bind, which is a problem merely in light of his being an agent.

We can thus describe this agent's problem in terms that abstract from the particulars of his case: He is irrationally conflicted. This description specifies the form of the agent's motivational set but not its content.

One might insist that the irrationality of being conflicted does too depend on the content of the agent's motivational set. Being conflicted is irrational, one might say, only because it frustrates the pursuit of a higher-order end that any agent must have, the end of attaining his lower-order ends. This end gives any agent reason to avoid having lower-order ends that cannot be jointly attained, and hence to avoid conflicts. But this way of stating the problem is misleading and consequently unpersuasive. Can't an agent's motivational set fail to include the higher-order motive that would give him the requisite end? What if an agent cares about his several ends but not about the master end of their joint attainment?

What's misleading about this statement of the problem is that, in mandating a higher-order end, it seems to be mandating a motive, as an element of the agent's motivational set; whereas an agent's motivational set is supposed to represent the contingent, individually variable input to his practical reasoning. This statement of the problem therefore invites the question why an agent must have a motive toward attaining his lower-order ends.

Yet if ends are conceived as variable between agents, because of arising from their individual motivational sets, then each agent must have something else – a project, it might be called – that isn't an end in this sense. An agent must have the project of coping with, or doing justice to, the reasons that issue from his motivational set (or from anywhere else, for that matter). The reasons there are for him to act define a practical

problem for him, and he must have the project of solving such problems, if he is to be a rational agent. This project isn't an end because it isn't just given to the rational agent by the contingent elements of his motivational set; it's a prerequisite for his being a rational agent, who can regard his motivational set (or anything else) as a source of reasons.

So even if we start from the assumption that reasons for acting must issue from the projects represented in the agent's motivational set, we end up at the realization that an agent must have at least one additional project, simply by virtue of being an agent – namely, the project of coping with the reasons that issue from his motivational set, a project that requires a motivational set that issues in reasons with which he can cope. Since the reasons that issue from deeply conflicting projects are extremely difficult to cope with, being conflicted is a hindrance to the project of practical reasoning itself.

I want to say that this hindrance to the project of practical reasoning renders the conflicted agent irrational. In so saying, however, I seem to be blaming a difficult situation on its victim. How can an agent be irrational for facing a difficult practical problem?

At this point, a contrast with theoretical reasoning might be helpful. In theoretical reasoning, we must cope with the various reasons for belief that confront us – the evidence, the arguments, our prior assumptions, and so on. Our task is to arrive at a belief that accommodates these reasons as well as possible, as if it were the solution to a set of simultaneous equations. Some sets of equations admit of an obvious solution and are therefore easy to solve; but other sets admit of no solutions, in which case we are obliged to discount some of our evidence, discard some of our assumptions, or otherwise adjust the set of reasons to be accommodated.

In theoretical reasoning, our task is to cope as best we can with whatever reasons the world serves up to us. A difficult theoretical problem is an inscrutability in the world, not an irrationality in ourselves. But in practical reasoning, the reasons with which we must cope, the simultaneous equations that we must solve, are served up by our personalities and circumstances, which are partly our own responsibility. The man in Williams's example didn't just wake up in a bind: he probably got himself into a bind by ignoring signs of trouble, shirking crucial choices, and making fateful compromises over time. More importantly, he can and ought to get himself out of his bind, though doing so will also take time. For he can and ought to resolve the conflicts in himself, by altering his motives or his circumstances, or both. So whereas a theorist with deeply conflicting evidence is merely unfortunate, an agent with deeply

conflicting projects may be rationally criticizable, insofar as he is responsible for getting into, and is in any case responsible for getting out of, his own deliberative difficulties.

One might object that the epistemic agent can also avoid deliberative difficulties, simply by closing his eyes to recalcitrant evidence or closing his ears to distracting hypotheses. But such maneuvers would defeat the purpose of theoretical reasoning, which is to arrive at the truth, or at least at the hypothesis that best accounts for the phenomena, where the truth and the phenomena are fixed by the way the world is. By contrast, the purpose of practical reasoning is not just to cope with reasons that are fixed by the agent's current motives and circumstances, since changing his motives or circumstances often remains one of the agent's options, a possible outcome of his practical reasoning. The epistemic agent's predicament is defined by what the world is like, and he must cope with that predicament, because he cannot change it. But the practical agent's predicament is defined by what his life is like, and one of the resolutions available to him is to change his life. The practical agent can therefore be held rationally responsible for getting himself into, or not getting himself out of, the wrong predicaments, predicaments that are wrong in the sense that they confront him with a problematic set of reasons.

How to Hold an Agent Responsible for His Reasons

Kant is committed to holding an agent responsible for the reasons that apply to him. This commitment appears in Kant's doctrine of willing the law.

An agent wills the law, according to Kant, when he wills his maxim in the form of a law for all rational agents; and his maxim is a principle of practical reason, specifying a proposed course of action and his reasons for taking it.[7] The agent wills that the principle of taking that course for those reasons be valid for any rational agent, as it would have to be in order to be rationally valid at all, even for him.[8] Thus, when the agent

[7] For this conception of maxims see, e.g., Korsgaard 1996a: 13: "Your maxim must contain your reason for action: it must say what you are going to do, and why."

[8] In many formulations, the Categorical Imperative appears to require only that the agent be *able* to will that his maxim become a universal law. But the best justification for requiring that he be able to will the universalization of his maxim is that he must actually will it, or at least regard himself as willing it. And this necessity is indeed asserted in Kant's Formula of Autonomy: "The principle of autonomy is . . . : to choose only in such a way

acts for reasons, he acts on the basis of considerations that he has willed to be valid as reasons, for himself or anyone else.

In the form enunciated by Kant and adopted by contemporary Kantians, the doctrine of willing the law does not fit the process that I have imagined as making an agent responsible for the reasons that apply to him. But that orthodox form of the doctrine is also flawed, in my opinion; and its flaws turn out to coincide with its differences from the process that I have imagined. In my opinion, my concessions to the critics of Kant turn out to be an improvement over the orthodoxy.

In the process that I imagined here, an agent is responsible for the reasons that apply to him insofar as he is responsible for his personality and his circumstances, which at any particular time determine the set of applicable reasons. But the agent's responsibility for his reasons, in my conception, does not involve the capacity to decide, at a particular moment, which reasons apply to him. His personality and his circumstances determine the set of applicable reasons in a systematic way that is not up to him;[9] and because he cannot change his personality or his circumstances on the spot, he cannot immediately change the reasons that apply to him, either. He is responsible for the reasons that apply to him only because his choices over time have shaped, and will continue to shape, the attitudes, traits, and circumstances that determine the set of applicable reasons.

By contrast, Kant's doctrine of willing the law seems to imply that an agent is in a position simply to will that particular considerations have validity as reasons – as if their rational force were up to him. This implication follows from a combination of passages, as follows. First, Kant defines the will as "the capacity to act *in accordance with the representation* of laws, that is, in accordance with principles"; he adds that "[s]ince *reason* is required for the derivation of actions from laws, the will is nothing other

that the maxims of your choice are also included as universal law in the same volition" (*Groundwork of the Metaphysics of Morals*, 47 [4: 440]). See also 45 [4: 437–8], "the basic principle, act on a maxim that at the same time contains in itself its own universal validity for every rational being," and 46 [4: 438–9], "act in accordance with the maxims of a member giving universal laws for a merely possible kingdom of ends." (Note that in the last quotation, what is qualified as merely possible is, not the agent's willing of his maxim as a law, but the kingdom of ends that would exist if the law were universally obeyed.)

[9] I haven't specified how the agent's motivational set determines the set of applicable reasons, because I disagree with Williams and other so-called internalists on this question. In particular, I don't believe that reasons applicable to an agent are dependent on his motivational set, as conceived by Williams, for their capacity to influence the agent's behavior.

than practical reason."[10] Kant subsequently asserts that "every rational being having a will" must exercise that will under the idea of freedom, because the will consists in practical reason and "[r]eason must regard itself as the author of its principles."[11] Thus, in any being with a will, practical reason must derive actions from laws of which it regards itself as the author; and those laws, as we have seen, are universalized principles of acting in particular ways for particular reasons. Any rational being must therefore purport to originate the principles expressing the validity of his own reasons for acting.

As I have said, I think that the orthodox Kantian doctrine of willing the law is flawed. I think that practical reason need not – indeed, cannot – regard itself as the author of its principles, because an agent cannot regard himself as originating the validity of his reasons for acting. I will now try to explain this flaw in the Kantian view and how it can be corrected.

The flaw in this conception of practical reason is that it cannot explain how an agent is guided by reasons for acting. The volition in which the agent wills the universal validity of his reasons is the same as the volition in which he wills his action, since his decision to act for those reasons "contains in itself its own universal validity for every rational being"[12] or is "also included as universal law in the same volition."[13] Because the agent's decision to act for reasons contains or includes his willing the validity of those reasons, it cannot be guided by any prior recognition of their validity. All that guides the agent's decision, according to Kant, is his recognition that he is not precluded from willing the universal validity of his reasons for acting. In framing his decision, the agent is not bound by any antecedently valid principles of practical reasoning other than the principle of framing his decisions as principles whose universal validity he can simultaneously will.[14]

Critics of Kant have long complained that when the Categorical Imperative is so understood, it does not constrain the agent's choices in the determinate way that morality constrains them, because it constrains their form but not their substance. I am not sympathetic to this complaint when it is directed against Kantianism as a moral theory, since I think that an important part of morality is precisely a constraint on the

[10] *Groundwork*, 24 [4: 412]; see also 36 [4: 427].
[11] *Groundwork*, 54 [4: 448].
[12] *Groundwork*, 46 [4: 438–9].
[13] *Groundwork*, 47 [4: 440].
[14] See *Groundwork*, 40 [4: 442]: "[T]he human being . . . is subject *only to laws given by himself but still universal* and . . . he is bound only to act in conformity with his own will, which, however, in accordance with nature's end is a will giving universal law."

form rather than the content of the will. But I am sympathetic to the complaint when it is directed against Kantianism as a theory of practical reason. In this capacity, the Categorical Imperative implies that the reasons for an agent's decision must be reasons whose validity is willed in that very decision. And a decision that wills the validity of its own reasons cannot be guided by a recognition of their validity.

The only guidance available for such a decision is the guidance of the Categorical Imperative itself, which rules out deciding to act for reasons whose universal validity cannot simultaneously be willed. Within this purely formal constraint, the agent can decide to act for any reasons that he thereby wills to be universally valid. But how can the agent regard his decision as being guided by reasons whose validity he regards as being conferred on them by that very decision? How can he actually *be* guided by reasons so regarded?

A deeper aspect of this problem is that the validity of reasons is not the sort of thing that we ordinarily conceive as being subject to the will at all. A reason for acting is a consideration that purports to justify acting, and a valid reason is a consideration that, if true, really does justify what it purports to. But to justify something is to show (at least prima facie) that it is just, in the archaic sense of being in accordance with a *jus*, or rule of correctness. Hence a consideration justifies an action by tending to show that it would be a correct thing to do. How can the agent decide whether a consideration tends to show that a particular action would be correct?

If Kant were to acquiesce in this manner of speaking, he would point out that we are puzzled by the notion of an agent's deciding the justificatory force of reasons only because we assume that they must exert that force in relation to antecedently fixed rules of correctness for action. Perhaps we are improperly assimilating the case of action to that of belief, in which reasons must exert their justificatory force in relation to the antecedently fixed rule that a belief is correct only if true. A believer is in no position to decide which considerations shall have validity as reasons for belief, because he cannot decide which considerations show a belief to be correct in relation to the rule of truth. But in the practical case, the agent can decide the validity of reasons, Kant would argue, because he decides the rules of correctness as well: the autonomous agent adopts his own rules of correctness for action, subject only to the proviso that he adopt them in universal form, avoiding any rules that he cannot thus universalize.

Thus, Kant would say, we were puzzled about willing the law only because we had too narrow a conception of this process, as a process of willing merely that particular considerations should count as showing

an action to be correct. That conception left us wondering how an agent could possibly decide the import that particular considerations would have for the correctness of an action. The answer is that we need a broader conception of willing the law, as a process of willing the rules of correctness themselves, and only thereby willing the validity of the associated reasons. The agent wills that actions of a particular kind shall be correct in circumstances of a particular kind – which amounts to willing that consideration of the circumstances shall tend to justify the actions.

Unfortunately, this clarification doesn't solve the problem. If one's actions are subject to no fixed rules of correctness other than a rule for willing what those rules of correctness shall be, then one cannot really place one's actions under rules of correctness after all, since one's latitude in willing those rules reveals that, when it comes to actions, anything goes. How can one make an action correct in the circumstances by willing it to be correct, given that one could equally have conferred such correctness on a different action? Willing the law now looks like an empty exercise in self-congratulation – a matter of ruling one's choice to be correct so that one can pat oneself on the back for choosing correctly.

Of course, Kant will respond that not *quite* anything goes when it comes to actions, because in willing rules under which his actions are correct, the agent is restricted to rules that he can will in universal form. Perhaps the rules that he adopts can be rules of correctness because they have been constrained by the master rule of universalization – that is, by the Categorical Imperative. Yet this response brings us back to the problem of empty formalism, regarded again as a problem in the Kantian conception of practical reason. When the agent doesn't know what to do, he looks for reasons to guide him; but all he finds, according to Kant, is a set of actions that (under some description) he could will to be universally correct in circumstances that (under some description) are similar to his. Even if we believe that this set would exclude any morally impermissible actions, we must doubt whether the agent can will distinctions of correctness among the remaining, permissible alternatives. Within the constraints of the Categorical Imperative, the agent appears to face an arbitrary choice among various universal rules, which would specify various actions as correct in light of the circumstances, variously considered, thereby constituting different considerations as reasons for taking different actions. Having decided to act under one of these rules, how can the agent regard it as conferring correctness on his action, or normative force on his reasons, given that he has simply adopted it from among various rules that would have constituted other permissible actions as correct, and other considerations as reasons?

Let me repeat that this problem is more difficult for a Kantian conception of practical reason than it is for Kantian ethics. The availability of many act descriptions that the agent could consistently incorporate into a universal law is not necessarily a problem for the Categorical Imperative in its capacity as a test of morality. The test of the Categorical Imperative applies to any description under which the agent proposes to act, and it yields an up-or-down verdict on the permissibility of acting under that description. Once the agent has discovered which of the available acts would be permissible, and which would not, he has completed the moral reflections required of him, according to Kant. But having sorted the available acts into the permissible and the impermissible, the agent has not yet completed his practical reasoning about which of the permissible acts to perform. He must still choose one of the permissible acts rather than the others, and he must choose it for reasons. (If he didn't have to choose on the basis of reasons, then his choice wouldn't have been constrained by the Categorical Imperative, since the necessity of choosing for reasons is what generates the need to universalize.) The problem is that the reasons favoring one permissible act or the other are reasons that the agent himself must will into validity as he chooses between them. So how can he look to the validity of these reasons as a basis on which to choose?

One might think that the problem is solved by the additional constraint of hypothetical imperatives, which require the agent to will adequate means to his ends. Yet hypothetical imperatives, too, are merely formal constraints that provide only minimal guidance. They require only that an agent either abandon an end or adopt adequate means to it; and in the latter case, only that he adopt some adequate means or other. Of course, many philosophers believe that such formal constraints exhaust the guidance available from practical reason, which does no more, in their view, than enforce consistency on an agent's choices. But this solution is not available to Kant, precisely because he regards every rational choice as adopting not just a particular action but a corresponding rule of correctness, which is more specific than the purely formal imperatives that constrain it. In choosing among the actions conducive to his ends, the agent must will a law conferring correctness on actions like his in circumstances like his, so that his choice is derived from a law of which he can regard himself as the author. (Otherwise, he wouldn't qualify as choosing at all, and unchosen behavior, not purporting to embody a rule, would not have to be universalizable.)[15]

[15] See the quotation from pp. 231–2 of Korsgaard's Tanner lectures (1996d), on p. 299.

Thus, the agent still appears to be engaged in an empty form of self-congratulation. The rule of correctness that ought to be the basis for a choice is only willed into force as the choice is being made. The problem is how the basis of a choice can be willed into being by the same volition as the choice itself.[16]

Korsgaard's Concessive Version of Kant

Christine Korsgaard grapples with these problems in her Tanner Lectures, where she offers her own model of willing the law.[17] I am going to argue that Korsgaard comes close to adopting what I call concessive Kantianism; but first I'll need to explain how the relevant criticisms bear on Korsgaard's version of Kantian ethics.

[16] I think that there are cases in which practical reasoning takes this puzzling form, but they aren't cases that support the Kantian conception. For they are clearly unsuited to be a model of practical reasoning in general.

When you are tempted to eat or drink too much, to work or exercise too little, to shirk a social obligation or make an undue imposition – in short, to indulge yourself in some way – you tend to look for aspects of the occasion that make it unusual, precisely so that you can endorse such self-indulgence only on similarly unusual occasions, while continuing to condemn it more generally. These circumstances needn't be ones that you antecedently regard as positive reasons for self-indulgence: they may be as trivial as the fact that it's Tuesday. Yet if you allow yourself, say, to overeat in light of the fact that it's Tuesday, then you seem to make its being Tuesday a reason for allowing yourself to overeat. You thus seem able to choose what shall count as a reason for your action on this occasion and others like it.

I think that when you preemptively excuse or rationalize an action by finding an acceptable principle for it, you may indeed be in the position of willing the law; and your principle may indeed have justifying force insofar as you accept it under the constraint of having to accept it in universal form. Not just anything goes when it comes to self-indulgence, only those things which you're willing to accept as going in general – including, perhaps, overeating on Tuesdays, but not overeating every day. And because you've confined yourself to what you're willing to accept as going in general, you seem to be justified in letting it go today. So you seem to have willed your action into being correct, and your circumstances into being reasons for it.

Although such cases exemplify the Kantian conception of willing the law, they don't lend that conception much support. For they are cases not so much of adopting reasons as of adopting excuses or pretexts – cases in which your principles at best permit you to do something but don't positively guide you to do it. Again, your reasoning in these cases may be an adequate model of moral reasoning, insofar as moral reasoning is just a matter of asking, "May I?" But practical reasoning is not in general permissive, not just a matter of asking, "May I?" It's a matter of asking what you should do from among the many things that you may. Practical reasoning must give you a positive basis for choosing among the many morally permissible actions, and cases in which you adopt principles for permitting or excusing self-indulgence cannot be a model for such reasoning.

[17] Korsgaard 1996d.

In Korsgaard's version of Kantian ethics, willing the law is a matter of adopting a self-conception, or "practical identity." Korsgaard derives the notion of practical identities from the phenomenology of reflective agency:[18]

> When you deliberate, it is as if there were something over and above all of your desires, something which is *you*, and which *chooses* which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of *yourself*. To identify with such a principle or way of choosing is to be, in St. Paul's famous phrase, a law to yourself.

Because an agent identifies with the principle that dictates his choice, the principle can actually be expressed as a self-conception:

> An agent might think of herself as a Citizen of the Kingdom of Ends. Or she might think of herself as someone's friend or lover, or as a member of a family or an ethnic group or a nation. She might think of herself as the steward of her own interests, and then she will be an egoist. Or she might think of herself as the slave of her passions, and then she will be a wanton. And how she thinks of herself will determine whether it is the law of the Kingdom of Ends, or the law of some smaller group, or the law of egoism, or the law of the wanton that will be the law that she is to herself. (101)

As we have seen, the law with which the agent identifies, by adopting one of these self-conceptions, is in fact a principle of choosing to act in particular ways under particular circumstances – circumstances that the law constitutes as reasons for acting in those ways. Thus:

> That you desire something is a reason for doing it from the perspective of the principle of self-love.... That Susan is in trouble is a reason for action from the perspective of Susan's friend; that the law requires it is a reason for action from the perspective of a citizen, and so forth. (243)

In sum, an agent adopts a principle that determines what counts as a reason, but he adopts that principle in the form of a conception of himself as someone's friend, or as a citizen of a nation, or whatever.

Korsgaard's lectures seem to equate practical identities with principles of choice. She appears to say that adopting the identity of Susan's friend just consists in identifying with particular principles of choice, such as the principle of helping Susan when she's in trouble. This view implies that insofar as the reason-giving import of Susan's troubles depends on whether the agent is her friend, it depends on whether the agent adopts

[18] Korsgaard 1996d: 100. All parenthetical references from here on are to that volume.

particular principles of choice, including the principle that explicitly specifies her troubles as reasons. The view therefore implies that the agent can decide whether Susan's troubles have reason-giving force for him, simply by deciding whether to adopt a principle conferring such force upon them.

Cohen's Objection and the Beginning of Korsgaard's Reply

In the symposium on Korsgaard's lectures, G. A. Cohen objects that a law adopted at will by the agent can just as easily be repealed by the agent and therefore fails to bind him in any meaningful way: "[A]lthough you may be bound by a law that you can change, the fact that you can change it diminishes the significance of the fact that you are bound by it. There's not much 'must' in a 'must' that you can readily get rid of."[19] To say that "there's not much 'must' " is to say that almost anything goes, and so Cohen's objection to Korsgaard resembles the one that I raised earlier against Kant. In either case, the objection is that being adopted at will would drain rules or laws of any significant normative force.

Korsgaard's answer to this objection shows that her conception of willing the law already differs from the conception that I have attributed to Kant. Her answer begins with a point that Cohen himself has acknowledged, "that even if I can change the law that I make for myself, I remain bound by it until I can change it" (234). What this point reveals is that in Korsgaard's model, an agent's decision is constrained, not only by the principle of choice that he wills in making that very decision, but also by principles that he has willed on previous occasions. And the latter principles are antecedently available to guide the agent's present decision, unlike the principle that he wills in making the decision itself.

An agent can thus be guided, in making his present decision, by reasons whose validity he has willed in the past. Each time he has made a choice in the past, he has willed and thus committed himself to a principle dictating similar choices in similar cases, including cases that he might encounter in the future. If he now encounters such a case, he will be bound by his former commitment to that principle of choice. In time, the agent will find himself encumbered with commitments to many principles, of which it is likely that some will apply to any particular case he may encounter; and he remains encumbered by those commitments until he revokes

[19] Cohen 1996: 170.

them. On any particular occasion, the relevant principles will constitute various considerations as reasons for him to act.

With this background in place, Korsgaard is now in a position to answer Cohen's objection. The objection, remember, was that even if a principle adopted by the agent is binding until revoked, it does not significantly bind him given that he is empowered to revoke it. Korsgaard's answer to the objection is this: "[I]f I am to be an agent, I cannot change my law without changing my mind, and I cannot change my mind without a reason." Hence "we cannot change our minds about just anything" (234). Korsgaard derives this answer from the results of an earlier discussion, in which she put the point as follows:

> If I am to regard *this* act, the one I do now, as the act of my *will*, I must at least make a claim to universality, a claim that the reason for which I act now will be valid on other occasions, or on occasions of this type.... Again, the form of the act of the will is general. The claim to generality, to universality, is essential to an act's being an act of the will.
>
> A couple of paragraphs ago I put into the objector's mouth the claim that when I make a decision I need not refer to any past or future acts of my will. But now we see that this turns out to be false, for according to the above argument it is the claim to universality that *gives* me a will, that makes my will distinguishable from the operation of desires and impulses in me. If I change my mind and my will every time I have a new impulse, then I don't really have an active mind or a will at all – I am just a kind of location where these impulses are at play. And that means that to *make up my mind* even now – to give myself a reason – I must conceive of my reason as an instance of some general type. Of course this is not to say that I cannot ever change my mind, but only to say that I must do it for a reason, and not at random. (231–2)

We'll need to spend a moment analyzing this passage in order to understand Korsgaard's conception of willing the law.

In this passage, Korsgaard observes that making up one's mind requires one to adopt some stable or settled practical stance, which must consist in more than an occurrent impulse. From this observation, she infers that making up one's mind requires one to have a general principle, which will embody one's made-up mind. The conclusion of this inference has both subjective and objective aspects. The subjective aspect is that in order to *view* oneself as making up one's mind, one must view oneself as instituting a stable practical stance; and one must attain this view by framing one's decision in the form of a general principle that purports to cover occasions beyond the present. But one can thus purport to make up one's mind without actually succeeding, since the purportedly stable stance that one has instituted may consist in a general principle that one

might instantly revoke at any time. Objectively speaking, then, making up one's mind requires not only that one's practical stance purport to cover future occasions but also that it really have some stability across the occasions that it purports to cover.

Korsgaard's argument continues from the latter, objective condition on making up one's mind. Before one can change one's mind, one must have succeeded in making it up one way or another, so that there is something for one to change; and in order to have made up one's mind one way or another, one must have arrived at a stance with some real stability. Hence one cannot change one's mind if it is unduly changeable, since what is unduly changeable does not amount to a made-up mind, to begin with. Changing one's mind entails becoming differently *minded*, which requires being antecedently minded in some determinate way, which in turn requires being resistant to undue change.

According to Korsgaard, this restriction on undue changes of mind restricts one to changes of mind for which one has a reason. What she says is that one must change one's mind "for a reason, not at random." Her thought appears to be that change at random is undue change, which made-up minds tend to resist, and that the opposite of change at random is change for a reason. If one changes one's mind at random, then it will not really have been made up, in the first place, and to that extent won't amount to a mind to be changed. But if one changes one's mind only for a reason, then one's mind, though proving to have been changeable, will not have been unduly so, and hence will really have been made up, after all. The requirement to have a reason for changing one's mind ensures that a change will amount to a transition between determinate ways of being minded rather than a dissolution of determinate mindedness altogether.

But now the problem of empty formalism reemerges. For if one's change of mind is not an undue change so long as it is based on a reason, then the ready availability of reasons will take the bite out of any restriction on changes of mind. In order to change his mind for a reason, the agent need only make the change under cover of a relevant principle, which will constitute some considerations as the requisite reasons. Of course, the agent may already be committed to principles about how and when to change his mind; but those principles will themselves be subject to reconsideration and revision, provided only that their revision be adopted under a yet further principle.

Let me clarify the problem by summarizing how it has arisen. Kant says that when an agent decides to take an action, he must will not just the action but also a relevant principle of correctness, which constitutes

particular considerations as reasons for taking the action. We worried that giving the agent this much latitude to bless his own actions as correct, and thus to constitute considerations as reasons for taking it, would undermine the very possibility of correctness in actions, or of normative force in reasons for acting. When Cohen expressed this worry, in response to Korsgaard's reconstruction of Kant, her answer was that the agent's latitude is significantly restricted by the blessings he has conferred on actions in the past, whereby he committed himself to principles of correctness, and hence to reasons for acting, to which he remains committed.

What worries us now is that the agent's latitude cannot be restricted by past commitments, precisely because it undermines those commitments as well. No rational commitment is so binding that it cannot be revoked for good reasons – that is, on the basis of considerations tending to show that revoking it would be correct. Hence the agent's latitude to confer correctness on actions, and thus to constitute considerations as reasons, cannot be restricted by past commitments, because it includes latitude to confer correctness on the act of revoking those commitments, and thus to constitute reasons for revoking them. Although the agent has committed himself by blessing his actions in the past, he can always revoke those commitments under cover of a new blessing. *The problem is that reasons are too easy for the agent to conjure up, and so the solution cannot be that once having conjured them up, he needs a reason for conjuring them away.*[20]

[20] The problem of empty formalism runs even deeper than Korsgaard's conception of practical reason: it runs all the way down to the agent's conception of his own agency. For in Korsgaard's version of Kant, principles of choice are constitutive not only of practical reasoning but of the agent himself. Korsgaard puts the point most clearly in a recent article:

> To conceive yourself as the cause of your actions is to identify with the principle of choice on which you act. A rational will is a self-conscious causality, and a self-conscious causality is aware of itself as a cause. To be aware of yourself as a cause is to identify yourself with something in the scenario that gives rise to the action, and this must be the principle of choice. . . . 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. . . . Self-conscious or rational agency, then, requires identification with the principle of choice on which you act. (1999: 26)

> Yet if the person casts himself as author of his actions by identifying with the principles that generate those actions, then how does he cast himself as the author of his principles, as Kant says he must? The answer would seem to be that he generates the principles of his actions from antecedent principles with which he also identifies. Yet this regress of principles cannot go on forever; and where it ultimately stops is at the Categorical Imperative, which is the principle simply of deriving things from principles – a purely formal principle, from which no particular substantive principles can be derived. [Note continues on p. 302.]

is limited. He can alter the range of available reasons only by adopting, shedding, or somehow modifying his practical identities, and this process takes time. Hence he cannot alter the available reasons on the spot: "Although I have just been suggesting that we do make an active contribution to our practical identities and the impulses that arise from them, it remains true that *at the moment of action these impulses are the incentives, the passively confronted material upon which the active will operates*" (240–1, emphasis added).

Finally, Korsgaard believes that the motives and principles associated with an agent's contingent practical identities can be genuinely normative even if they are ultimately in conflict with the Categorical Imperative. Discussing the case (introduced by Cohen) of a man whose practical identities include that of a Mafioso, Korsgaard says:

> It would be intellectually tidy, and no doubt spare me trouble from critics, if I... said that only those obligations consistent with morality are "real" or in Cohen's phrase "genuine." Then I could say that it seems to the Mafioso as if he had an obligation to be strong and in his sense honour-bound, but actually he does not. I could say that there's no obligation here, only the sense of obligation: no normativity, only the psychic appearance of it.... But I am not comfortable with this easy way out, for a reason related to one of Cohen's own points – that there is a real sense in which you are bound by a law you make for yourself until you make another.... There is a sense in which these obligations are real – not just psychologically but normatively. And this is because it is the endorsement... that does the normative work. (257)

The endorsement that "does the normative work" of obligating the mobster to his mob is embodied in the principles that make up his identity as a Mafioso – principles of perfect loyalty to the mob and perfect ruthlessness to outsiders. And these principles underlie not only the mobster's obligations, when he is tempted to be less than completely loyal or completely ruthless, but also his reasons for being loyal or ruthless on occasions when he isn't tempted to be otherwise. If these principles can lend normative force to the obligations, then they must also be able to lend normative force to the associated reasons.[24]

[24] Korsgaard describes the relation between obligations and reasons as follows:

> To make a law for yourself... is at the same time to give expression to a practical conception of your identity. Practical conceptions of our identity determine which of our impulses will count as reasons. And to the extent that we cannot act against them without losing our sense that our lives are worth living and our actions are worth undertaking, they can obligate us. (129)

Korsgaard believes that the normative force of these reasons, like that of the associated obligations, can ultimately be undermined by the mobster's more fundamental identity as a human being who must act for reasons, the identity that consists in his commitment to the Categorical Imperative: "If Cohen's Mafioso attempted to answer the question why it matters that he should be strong and in his sense honour-bound even when he was tempted not to, he would find that its mattering depends on the value of his humanity, and if my other arguments go through, he would find that that commits him to the value of humanity in general, and so to giving up his role as a Mafioso" (256). But Korsgaard does not say that the existence of this latent conflict between the mobster's commitment to humanity and his commitment to the role of a mobster already undermines the normative force of the latter commitment, even before the conflict is discovered and the latter commitment revoked. On the contrary, she says that the latter commitment gives rise to genuinely normative obligations.

The resulting view severely constrains an agent's latitude in constituting and reconstituting reasons for acting. Reasons for him to act must consist in impulses endorsed by principles; his impulses are "passively confronted material" that he cannot change at the moment of action; and his principles can be revised only on the basis of reasons, which themselves require passively confronted impulses and/or conflicting principles to dictate the revision. Hence the agent cannot simply conjure up reasons for acting, or reasons for revising his principles, since he is confined in both instances to reasons based either on impulses already available in his motivational set or principles already available among his practical identities. Changing the set of available reasons therefore requires substantive psychological change, which the agent cannot effect at will.

In sum, we can no longer object that anything goes for a rational agent. His psychological makeup now provides substantive constraints on his practical reasoning, and so his practical reasoning is no longer an empty form. At the same time, however, the agent's practical reasoning is no longer guaranteed to encounter a set of reasons that weighs in favor of moral action. That's why I think that Korsgaard's view has become concessive.

Why Korsgaard's Version of Kantianism Is Concessive

Consider again the mobster introduced by Cohen. This man may have inherited the practical identity of a mobster from his family, or adopted

that identity on his own, or acquired it through some combination of these processes. In any case, his acquisition of that identity will have entailed the acquisition of associated desires and impulses, such as the desire to kill anyone who threatens the interests of the mob. When anyone threatens those interests, the desire to kill him will unavoidably arise as "passively confronted material on which [the mobster's] will operates." And his identity as a mobster will include principles endorsing such desires as reasons for acting. As endorsed by those principles, his murderous desires will have genuine normative force as reasons for the mobster to act. He will therefore have genuine reasons for committing murder.

To be sure, the mobster also has countervailing reasons, based in his fundamental identity as a human being, as expressed in the Categorical Imperative. But these reasons weigh against acts of murder only indirectly, by committing him to "giving up his role as a Mafioso." They are reasons for him to revoke his commitment to that more particular identity, which turns out to conflict with his underlying identity as a human being, and so they are reasons for him to become someone who no longer has reasons for committing murder. The mobster is irrational to commit murder, not because he doesn't have reasons for committing such an act, but rather because he has reasons against being the sort of person who has those reasons.

One might think that the mobster's fundamental commitment to his humanity, as expressed in the Categorical Imperative, militates directly against acts of murder, thus overriding the reasons generated by his identity as a mobster. The Categorical Imperative does militate directly against particular immoral acts in Kant's own version of the theory. Unfortunately, Korsgaard's version of the theory has eliminated the mechanism by which it militates against those acts.

In Kant's theory, the Categorical Imperative rules out particular acts only by ruling out the volitions behind them; and it rules out those volitions by requiring every volition to include not just the description of an action but also a universalized maxim of acting under that description – what Korsgaard calls a principle of choice. In order to commit a murder, the agent must will not just the particular killing but, at the same time, a principle of killing. And willing such a principle turns out to involve a contradiction. Because Kant's theory requires the agent to will the particular act only in conjunction with a principle, the contradiction involved in willing such a principle stands in the way of willing the act.

In Korsgaard's version of the theory, however, the agent may already be committed to the relevant principle by virtue of having adopted it

earlier and not repealed it since. In that case, there would seem to be no need for him to will the principle afresh in acting on it again. Indeed, he would seem to be in no position to will the principle any more, given that he is already committed to it: He can no longer will it to be a law for him, because it already is one, whether he likes it or not. And if his volition to act need not encompass his principle as well, then the contradiction involved in willing the principle cannot pose any rational obstacle to the act.

Imagine that Kant himself wrote in Korsgaard's language of self-constitution. In that case, Kant would say that the Categorical Imperative requires that, in choosing to kill, the agent adopt the identity of a killer, by adopting a general principle of killing. Kant would add that adopting the identity of a killer entails a contradiction that consequently stands in the way of choosing to kill. But Korsgaard's view is that the agent may already *have* the identity of a killer, and so the contradiction that would be involved in adopting that identity no longer stands in his way.

The result of retroactively imposing Korsgaard's terminology on Kant himself is a theory of radical self-constitution: with every act, the agent re-adopts the relevant identities all over again, reconstituting himself as a killer with every killing, as a friend with every act of friendship, and so on. This theory raises the problem of empty formalism precisely because it has the agent reinventing himself from the ground up with every choice. Because the agent can reinvent himself, he can rewrite the set of available reasons, and so almost anything goes. Korsgaard solves the problem of empty formalism by restricting the scope of the self-constitution that accompanies a particular choice: the agent approaches each choice with antecedently fixed identities, which he can revise only within constraints fixed, in part, by those identities themselves. Unfortunately, this solution to the problem of empty formalism removes the mechanism by which the Categorical Imperative militates against individual actions, since the Imperative militates against actions only by requiring them to include bits of radical self-constitution that would be contradictory. I therefore suspect that Korsgaard cannot avoid the concessive version of Kantian theory, in which moral considerations do not necessarily provide sufficient reasons against immoral acts.

Conclusion

This concessive version of Kantian theory has the strength of entailing weaker consequences than the orthodox version. It doesn't imply that every agent, on every occasion, has reasons for acting that on balance

forbid committing murder. It concedes that the first-order reasons available to a particular agent on a particular occasion – the reasons for choosing one action over another – may on balance favor his committing murder. It merely adds that someone who finds himself with such a set of first-order reasons will also have higher-order reasons for changing the reasons available to him, by changing himself. And then it adds the further concession that changing himself will take time.

Korsgaard puts the point like this:

> I am certainly not suggesting that the *rest of us* should encourage the Mafioso to stick to his code of strength and honour and manfully resist any wanton urges to tenderness or forgiveness that threaten to trip him up. The rest of us should be trying to get him to the place where he can see that he can't see his way to this kind of life anymore. (257)

In order to maneuver the Mafioso out from under the force of reasons for committing murder, then, we would have to "get him to a place" from which he could see something that he can't currently see from the place he's in, at the moment of pulling the trigger. Indeed, we'd have to get him to a place where he could turn around and see that he couldn't find his way back, a place that would therefore have to be far removed, in the space of reasons, from the place he currently occupies. Such changes of perspective cannot be brought about on the spot, when push has already come to shove, or shove to shoot.

Because this concessive Kantianism entails weaker consequences than the orthodoxy, it is harder to attack and easier to defend. Although Korsgaard suggests that it will evoke "trouble from critics," it will in fact disarm the traditional critics of Kant, who can no longer adduce the usual "hard cases" as counterexamples. The existence of hardened immoralists, who have no first-order reasons for doing the moral thing on some occasions, is perfectly compatible with the concessive version of Kantian theory. Critics will therefore have to go further afield for their counterexamples.

Here is a problem, though. If Korsgaard's version of the theory doesn't bring the Categorical Imperative to bear on particular murders, but only against being a person who has reason to commit them, then maybe it doesn't condemn murders as immoral; maybe all that it condemns as immoral is a willingness to acquire, or an unwillingness to shed, the identity of a murderer. This leniency may be implicit in Korsgaard's description of how "the rest of us" should regard the Mafioso. When she says that we "should be trying to get him to the place where he can see that he

can't see his way to this kind of life anymore," perhaps she means that we should blame him for going astray in life but not for pulling the trigger here and now.

Note that *this* moral theory wouldn't fit my model of concessive Kantianism, since it would preserve the equivalence between morality and rationality in action. The mobster would have sufficient reason to commit murders, but the murders that he committed would not in themselves be immoral; what's immoral would be his acquiring or failing to shed the identity that provides his reason for murdering, and this act or omission would indeed be contrary to the balance of reasons for him to act. So the fact would remain, as stated in proposition (ii), that wrongdoing entails irrationality in action; the extension of the term "wrongdoing" would merely have shrunk, to include primarily acts of self-constitution rather than garden-variety, first-order acts.

But surely Korsgaard's concession to the normativity of the Mafioso's commitments is not meant to imply that they are normative in the moral sense. I assume that Korsgaard believes the mobster's killings to be morally wrong, even though he has normatively potent reasons, and perhaps even obligations, to commit them. I assume that when she recommends coaxing the Mafioso into a different "place," she doesn't mean that this therapeutic approach should preempt moral condemnation of his actions. The therapeutic approach is the only way to reason with the mobster, given that his existing identities support only a defective set of reasons; but gently reasoning him out of his identity as a mobster is meant to be compatible, I assume, with uncompromising condemnation of what that identity leads him to do.

I have no idea whether these suggestions capture Korsgaard's intentions, but I think that they capture what is plausible in her treatment of the case. And they imply that Korsgaard has brought us to a version of Kantian ethics in which morality and rationality really do come apart. In this version of Kantianism, which really is concessive, what rationality recommends on a particular occasion is that an agent do what he has the strongest reasons for doing, even if those reasons arise from an identity that's irrational for him to have. But morality requires an agent *not* to do things for reasons that arise from irrational identities: morality requires him to act only on reasons that he could rationally have.

If this theory is right, then what becomes of reasons for being moral?

The theory insists that every agent has reasons for being a sort of person who has reasons for acting morally. In this sense, it insists on reasons for *being* moral. And the reasons for being moral, in this sense, are the

ones defined by Korsgaard's version of contradictions in the will: they are reasons that arise from underlying conflicts between immoral identities and the Categorical Imperative, which expresses the fundamental identity of a person. Because of these conflicts, being an immoral person is an irrational way of being a person, and so it isn't a way that any person could rationally choose to be, or to continue being. Therein lies the contradiction in the will of an immoral person, according to concessive Kantianism.

But concessive Kantianism doesn't insist on reasons for *acting* morally – not, at least, for agents who have failed to heed their reasons for *being* moral. If an agent has overlooked or tolerated the contradictions involved in having an immoral identity, he may then have insufficient reason for acting morally, according to this theory. This much the theory concedes, not only to the critic of Kant, but also to the immoralist.

Even so, the theory has one remaining resource for softening this concession. It can point out that acting morally represents, as it were, a higher rationality – a rationality of acting on the reasons of one's ideally rational self rather than one's actual selves. What morality requires one to do may not be what one actually has reason for doing, but it is what one could have reason for doing if one had a rational set of reasons.[25] And the same cannot be said for immoral actions. To be sure, it's rational to act on the reasons one actually has, even if they favor acting immorally. But to act instead on reasons that it would be rational to have is not exactly irrational: it is rather extrarational, above and beyond the call of practical reason.[26]

Consider again how the Mafioso might find his way out of the bind created by his immoral identity. As Korsgaard points out, he might reason his way out, but only by way of a long and subtle train of reasoning, which is unavailable in the heat of the moment. Yet even in the heat of the moment, the mobster might simply step out of his bind: The scales might fall from his eyes, and he might drop his gun and walk away, never to return to his life of crime. (Of course, this sudden change of practical identities might not be accepted by his former associates without help

[25] Why do I say "what one *could* have reason for doing. . . ."? The reason is that if one has an irrational set of reasons, then there is no particular set of reasons that one would necessarily have if one had a rational set instead. There is no particular identity that the Mafioso would necessarily adopt instead of his identity as a mobster. There are many ways for him to be moral, and morality requires only that he adopt one of them.

[26] What I'm suggesting, then, is that although the moral act is not always rationally required, under concessive Kantianism, it is at least rationally supererogatory.

from the Witness Protection Program.) In the latter case, I would say, the mobster would not be acting on the balance of reasons that were currently available to him. Rather, he would be rejecting some of the reasons available to him, thereby reconstituting his current set of reasons.

As I have said, the act of reconstituting his current set of reasons is not supported by the overall balance of reasons in that set. Shedding his identity as a mobster would be a betrayal of the mob and hence of the commitments fundamental to that identity. Hence it is not a rational step for the agent to take, all things considered. But the act of reconstituting his set of reasons is indeed supported by a crucial subset thereof – namely, the reasons arising from his underlying identity as a rational human being. And the agent can act on that subset of reasons while holding the others in abeyance; for he can think of himself merely as a human being, reflecting with critical detachment on his more specific identities. Thus, he can tentatively suspend his identity as a mobster for the sake of considering whether to reject it altogether.

Even this tentative suspension of an identity would not be rational for the agent, all things considered. His commitments to the mob strongly militate against even toying with the idea of betrayal. But he can still toy with the idea, albeit irrationally. Indeed, he can *literally* toy with it, by playing or pretending for a moment that he isn't committed to the mob. He can imagine himself to be only a part or aspect of everything that he is, so as to make believe that he is deciding from scratch what to be.[27]

Let me repeat that toying with an idea in this fashion would be an irrational process, since it would require the agent to pretend that he didn't have commitments and reasons that he actually has. But this irrational process would enable the agent to become a more rational person, who wasn't caught in a bind of conflicting reasons. The process would therefore constitute an irrational leap to a greater rationality – a leap of faith in the possibility of being more rational. Kant might call it a leap of faith in oneself as a person.

[27] For further discussion of this process, see Chapter 13 in the present volume. I discuss another instance of the same process in Chapter 14 in the present volume.