Does the Categorical Imperative Give Rise to a Contradiction in the Will?

Elijah Millgram

The *Brave New World*-style utilitarian dystopia is a familiar feature of the cultural landscape; Kantian dystopias are harder to come by, perhaps because, until Rawls, Kantian morality presented itself as a primarily personal rather than political program. This asymmetry is peculiar for formal reasons, because one phase of the deliberative process on which Kant insists is to ask what the world at large would be like if everyone did whatever it is one is thinking of doing. I do not propose to write a Kantian *Brave New World* myself, but I am going to ask, of what these days is called "the CI-procedure," what would happen if everybody followed it. I will argue that if the CI-procedure works as advertised, it exposes a practical incoherence in the commitment to having it govern one's actions: in the Kantian vocabulary that goes with the territory, that the Categorical Imperative gives rise to a contradiction in the will. (Less formally, that it is self-refuting.)

My target will be a recently influential interpretation of Kant, due primarily to John Rawls and a number of his students, most prominently Onora O'Neill, Christine Korsgaard, and Barbara Herman, a group I will for convenience refer to as the New Kantians.[1] Although it does draw on earlier interpretive work, this body of writing is relatively self-contained, and manageable in a way that the Kant literature as a whole no longer is. I don't myself wish to take a stand on whether the New Kantian reading is exegetically correct; it suffices for present purposes that it has proven itself interesting, plausible, and powerful enough to have moved Kantian moral philosophy back from the marginalized position it occupied a little over a quarter-century ago to the center of contemporary ethics.

I will begin by rehearsing the CI-procedure and the theory that accompanies it; the reader is warned that the setup will take more time than is usual in papers of this kind. Kant himself used the label 'Categorical Imperative' to mark three ideas that he thought were at bottom the same: the practical priority of universalizability, of respect for persons, and of autonomy. They are, however, at any rate on the surface, rather different, and in order to sidestep the issue of whether the different versions of the Categorical Imperative are in fact equivalent, I

will be focusing only on the first, namely, on Kant's insistence that one act only according to maxims of which one can at the same time will that they become universal laws.[2] So the first task on our agenda will be describing how the New Kantians reconstruct that demand.

1.

The first formulation of the Categorical Imperative supplies a test for the permissibility of a proposed action. The New Kantians render this test procedurally, and have come to call it the *CI-procedure*.[3] When it occurs to you to do something, you are to

1. Identify the *maxim* of the action.

> The maxim is the "subjective principle of volition" (G 400 n.) that underlies the action. It captures your understanding of the action and of why you are proposing to perform it. The New Kantian account depicts maxims as having something like a logical form:

> In circumstances *C*, to do *A*, because *P.*

> Here *A* is a description of the type of action; *C* specifies the occasions that are to trigger actions of type *A*; and *P* specifies the point of the action.[4]

2. Consider the maxim universalized, that is, imagine a world (a "perturbed social world," in Rawls's phrase) in which everyone in your circumstances (that is, in circumstances that share with the one at hand the features you understand to be relevant) does what you are proposing to do.

> If you *can't* do this—if such a world is literally inconceivable— or if the intention expressed in your maxim is bound to be frustrated in such a world, then your maxim fails the *contradiction in conception* test.[5]

> Each of the elements of a maxim plays a role in the New Kantian reconstruction of the CI-procedure. *C* and *A* are the clauses of the rule that is always acted on in the perturbed social world. By specifying the intended achievement, *P* gives content to the notion of a frustrated plan of action; if *P* is frustrated by executing the plan, then the plan is self-frustrating.

3. Ask whether there are intentions that you are bound to have simply in virtue of being a human agent, but that cannot be successfully executed if your maxim is universalized.

   If there are, your maxim fails the *contradiction in the will* test.

4. If your maxim passes both tests, you may go ahead and perform the action; if it does not, acting on your maxim is prohibited.

What this comes to is best made clear by example, but before I get to that, a couple of clarificatory remarks. First, although I have described the procedure as something you pause to execute before going ahead with an action you have in mind, the Kantian requirement is of course not that you stand around muttering to yourself before you do anything. The procedure is meant as a rational reconstruction of the deliberative background to a decision properly arrived at, in pretty much the way that Aristotelian practical syllogisms are meant to reconstruct a somewhat different kind of deliberative background to action (Rawls 2000, 218).

Second, the New Kantians understand the point of the CI-procedure to be practical consistency.[6] The idea is that self-frustrating plans of action are the analogues, in practical reasoning, of the kind of incoherence that contradictory beliefs amount to in theoretical reasoning. Uncontroversial models for such self-frustration can be found in means-end incoherence, as when you decide to go to New York, but tear up the ticket that would get you there, or (to borrow an example from Garry Trudeau (1996, 104)) when someone decides he should have gone to medical school, but that dropping out of high school was definitely the right choice. You cannot coherently intend a self-frustrating plan of action, and the CI-procedure is presented as a way of checking whether what you are proposing is something that you can coherently intend. It is not (as such illustrious readers as John Stuart Mill have believed) a way of checking whether the results of everybody acting as you propose would be to your liking. I will defer to another occasion the question of *why* intentions that fail the CI-procedure are supposed to be on a par with self-frustrating plans; for now, we need to bear in mind that courses of action with genuinely awful consequences can pass the CI-procedure, which is to say that it is a deeply nonconsequentialist way of thinking about what to do. The question is: "*Can* you (not: *do* you) will that everybody do as you are proposing to do yourself?"

Kant illustrates his proposal with four examples, and because it is important for the subsequent argument to have the moves clearly laid out, I will walk through four examples as well, construing them as the New Kantians do. However, Kant's treatment of suicide is hard to bring into line with the New Kantian reading (and in fact it is not easy to see how suicide is an appropriate example of Kant's claims on just about anybody's reading). Because we do not want to skimp on exemplary contradication-in-the-will arguments before proceeding to develop our own, I will substitute an alternative, the recent New Kantian argument against violence.

Lying. Suppose your maxim is: to lie about whether you can and will pay back your creditor, whenever you need a loan that you're not in a position to repay—the point of your action being, of course, to get the money. If we think about a world in which everybody lied in these circumstances, we realize that in that world your plan of action could not possibly be effective; no creditor would believe you, and the lie would not work. Willing both a world in which everybody does as you do, and that your lie be effective, is something very much like adopting a self-frustrating plan (though, again, we haven't said why you have to be committed to both sides of the "plan"). Therefore, lies of this type are strictly impermissible. Restrictions generated by failure at the contradiction-in-conception stage are "perfect duties"; there are no exceptions to the prohibition on acting on this maxim.[7]

Before moving on to the next example, I want, for reasons that will be apparent in the sequel, to give this one a little more discussion than it usually gets. Kant's argument seems to depend on an empirical premise—that a *practice* of lying will undercut its own effectiveness—that is obviously often false.[8] Airlines, for instance, routinely publish schedules that they know they cannot meet; their maxim is, roughly, "When it will allow us to utilize our capital more efficiently, we will announce schedules we can't possibly stick to, in order to increase revenue." But (as you can confirm by looking around, the next time you are in an airport) passengers have not ceased to believe the schedules: they are surprised and upset when their flights are delayed, they have plans made around their flights' scheduled arrival times, and they are completely unprepared when they turn out to have missed their connections.[9] As a matter of psychological fact, people simply do not behave in the way that Kant's argument says they do. There are two ways we might approach this problem, and I just want to indicate what the Kantian's choice is. He can treat the premise as empirical, and

reshape his moral arguments around whatever the psychological facts turn out to be. Or, and I am myself inclined to think that this option is the more Kantian in spirit, he can treat it as prescriptive: Kant is on about how people are *supposed* to reason, and not about how they actually do. A rational agent *should* stop believing when it is obvious that there is a practice of lying; and we are to draw our moral conclusions on that basis rather than the empirical one.[10]

Mutual Aid. Suppose one of your maxims is: when someone needs a hand, not to help out, because you have other things to do with your time. Now, a world in which no one helps anyone else is (at first glance) conceivable.[11] But such a world would frustrate, not perhaps the intention expressed in the maxim, but a practical commitment you are bound to have to your own agency. In the Kantian picture, to be a rational agent is to be a creature that deliberates about, settles on, and then pursues its own ends. You have no way of knowing, now, what ends you will settle on down the road.[12] But you do know, humans being what they are, that if people are going to be agents worth the name, they will adopt projects that they cannot manage entirely on their own, and what is more, that the need for assistance crops up frequently enough in situations where there are no formal arrangements for assuring and compensating it. I decide to move my kitchen table, and so I have to get someone to lift the other end. I am lost, and since what I am is *lost*, the person from whom I have to ask directions is someone I do not already have anything like a contractual relationship with: I will have to ask a stranger for a favor. A world in which no one helps out will be a world in which the pursuit of your ends will predictably (often but not necessarily always) run aground. To will such a world is to will a world in which your agency is routinely frustrated, and your stake in your own agency is such that this would amount to a contradiction in your will. The "maxim of indifference" must be rejected.

Development of Talents. 'Talent' is here a misleadingly highfalutin word for the specialized skills of one kind or another that pursuing your ends is almost bound to require. There are many projects that you might reasonably adopt that would require driving to bring them off, and so being able to drive counts as a "talent" in the appropriate sense. It is an empirical but unavoidable fact about human agency that you yourself cannot be expected to have all the necessary skills; social arrangements built around the division of labor make it reasonable to expect that suitable resources of this kind will be available at the appropriate stages of your project. A world in which no one takes opportuni-

ties to develop his talents is, at first glance, anyway, conceivable.[13] But in such a world, those capacities are unavailable to you as resources, and your stake in your own agency—in the New Kantian reading, a kind of necessary end—is bound to be frustrated. Therefore, you cannot will a world in which the maxim we are considering is universalized, and the maxim must be rejected.[14]

Because Kant holds that the contradiction in the last two cases is in the will, rather than in conception—in the New Kantian reading, that the plan of action is not itself self-frustrating, but that when it is taken together with other ends that you necessarily have, these are jointly mutually frustrating—he holds that it gives rise to "imperfect duties." You obviously cannot develop all the useful specialized skills and capacities; you obviously cannot help out every time someone needs it. The argument is supposed to show that you have to help *sometimes*, and that you have to develop *some* skill set; but just when to help and when not to is left up to you, and which talents to develop is also left up to you.[15]

Violence.[16] Suppose your maxim is to take violent means, and in particular to kill, when that will promote your interests and projects. Being a victim of violence generally, and being killed in particular, tends to interrupt one's plan of action. (There are odd exceptions—for instance, when being bludgeoned to death is actually part of your plan.)[17] In a world in which everyone acted on this maxim, your agency in general would be ineffective because interrupted, and whatever project you are currently pursuing via the maxim would be interruptable.[18] That your current project be interrupted (or anyway interruptable by anyone else in pursuit of their ends), or that your agency generally be aborted or abortive, is not something you can coherently will. The violence-endorsing maxim must be rejected.

Notice a few features of New Kantian applications of the CI-procedure that turn on contradiction in the will (that is, of the latter three examples). First, they exploit deep facts about specifically human agency—about the range of ends that it is reasonable to expect humans to adopt, about the inability of human beings to do everything for themselves, or to acquire all the skills their projects are likely to need, and about the vulnerability of human agency to violent interruption. These are facts about people, not necessary features of agency. There might be creatures of whom none of this was true, creatures whose individual capabilities, skill sets, and robustness extended well beyond the range of their reasonably adopted ends, rather than falling short of it. If we were characters from *Road Runner*, the New Kantian argument

against violence would not go through: when Wile E. Coyote is crushed under a falling rock, he emerges slightly crumpled, but still ready to order the next Acme product. Second, the argument does not suppose that *all* conceivable projects, or even the projects you have actually undertaken, will be aborted if others do not come to your aid, etc.[19] If the argument is to work, the requirement must be rather that the world of the universalized maxim pose *enough* of an impediment to the range of projects that a rational deliberator—a creature that sets its own ends—might well adopt. And third, what drives the argument in each case is your stake in your own effective agency, which makes it impossible for you to coherently endorse a commitment to arrangements that would very broadly undermine it.

2.

New Kantians are committed to actions coming in all sizes and levels of abstraction. And refrainings and omissions can count as actions, provided that they are governed by one's intent in the same way that more obvious actions are.[20] It follows that governing one's activity by the CI-procedure—that is, not performing an action if its maxim does not pass the CI-procedure—is, anyway when one is "acting from the moral law," itself an action or plan of action. Like other actions, it can be the dictate of many possible maxims, but a maxim in line with the spirit of (New) Kantian moral theory would be:

> When I am making up my mind what to do, I will act only on maxims that pass the CI-procedure, so as to make (morally or rationally) permissible decisions.

Call this the *CI-maxim*; Kantians are committed to requiring of agents that something along the lines of the CI-maxim capture their volitional stance.[21]

New Kantian moral theory imposes the CI-procedure as a test that all of one's maxims must pass. The CI-maxim, which expresses the willingness to adopt this constraint, is itself a maxim. Therefore, proceeding on the basis of the CI-maxim must be contingent on its passing the CI-procedure.[22]

The point here is *not* that the CI-procedure has to show that the CI-procedure is mandatory; to establish that, one would proceed by testing a maxim containing the clause, "… not to act on the CI …," which is not what I propose to do. The thought is rather that because the CI-procedure tests the maxims you bring to it for practical consistency,

you do not want a maxim to fail—even if you think of it as foundational.[23] (Compare: in a foundationalist epistemological structure, we may exempt its foundational elements from having justifications, but we would still have cause for complaint if their contents turned out to have the form $p \wedge \neg p$.)

I will now walk through the application of the CI-procedure to the CI-maxim; as previously announced, I will be attempting to show that the CI-maxim fails at the contradiction-in-the-will stage of the procedure. We have just completed the initial step of the procedure, that is, identifying the maxim that it is going to be run on. The next step is to represent the perturbed social world in which the maxim is universalized. In the case of the CI-maxim, this is a world in which all agents treat the CI-procedure as a constraint on their actions. That is, they act only when they could will the maxim of their action to be universalized.

In willing that everyone always act in the way you are proposing, "as if the maxim of your action were to become through your will a universal law of nature" (G 421), you are willing that you act yourself in the way you are proposing, as though governed by "a universal law of nature"; after all, "everyone" includes yourself. That is, accepting the CI-maxim involves, in your own case, understanding your actions as governed by, or as the deliverances of, lawlike policies. (I am not just bearing down hard on something Kant happened to say; I will in due course argue that the CI-procedure can work only if its inputs are lawlike policies.) For a policy to be lawlike is for it to have no exceptions. So the perturbed social world is one in which, when anyone acts, he understands himself to be acting from a universal and so exceptionless policy that governs his action.

Of course we are not supposing that agents in the perturbed social world act on the *same* policies (except for the policy expressed by the CI-maxim). Agents in that world do will that others act on the same policies as themselves. But "willing" has, in one's own case, consequences it does not have in others'. When I will that I act from a policy or according to a law, that has the effect of committing me to act (it amounts to an intention to act) in accordance with the policy. But when I "will" that someone *else* act in accord with my policy, that need have no consequences for whether he does: most other people are not subject to my will. I also do not think we need to imagine that agents do not change their policies from time to time. They may decide that a previous policy was mistaken, and is to be replaced. What *is* required is that in so doing they understand themselves to have discarded one uni-

versal, exceptionless policy and to have replaced it with another. Kant has sometimes been accused of having a moral theory that generates exceptionless rules (that everybody has to abide by); that was a misperception of a different feature of the view, which is that it *operates on* exceptionless rules (but different ones for different agents at different times).[24]

What this means is that the perturbed social world of the universalized CI-maxim is one in which requests for exceptions to people's policies will be uniformly denied. We will have to proceed carefully here, making sure as the argument develops that we know just what this means. But meantime, notice that we are still on Kant's home turf. Kant diagnoses the immoral person as wanting to make an exception for himself (G 424); but if making an exception for oneself on one's own behalf is illegitimate, surely demanding that others make an exception for one must be illegitimate, too.

Now I want to advance the following claim: successful agency requires exceptions from others' policies, in just the way that successful agency requires assistance from others, in just the way that it requires immunity from violence, and in just the way that it requires the availability of a rich set of skills all of which one cannot have acquired oneself.[25] If this is correct (and if the Kantian model arguments for mutual aid and the like work as advertised), then, by parity of argument, the CI-maxim gives rise to a contradiction in the will: one's stake in one's own agency is such that one cannot endorse having it undermined by being deprived of the exceptions that are its precondition. And if that is in turn correct, then it is forbidden to act on the CI-maxim, and Kantian moral theory is (at least in its New Kantian rendition, and stating the conclusion informally) self-refuting.

Even though the analogous claims in the New Kantian model arguments are not taken to need support, I am going to argue for this one. I will proceed first by giving examples of the kind of case I have in mind. Then I will give an argument meant to explain why cases of this kind are common enough to be an empirical but deep fact about human agency. Finally, I will take up the New Kantian's stock objection to treating the cases as my argument requires—that is, to treating them as bona fide exceptions—and give two counterarguments to it.

Parking in Milan: I'll change a real-life story around a little bit to get my first, very small-scale example. When I was visiting friends in Milan, I needed to run my bags up to their apartment before returning the rental car; I found myself on a one-way and heavily-trafficked street,

with no on-street parking, and a barrier to keep cars off the sidewalk, which meant that I needed a small exception (the kind that involves blinking hazard lights) to the rule that governed the building's parking lot: only cars with permits. As it happened, the guard staffing the lot wasn't handing out exceptions, and my plans for the day had to be rewritten on the fly.

Trouble in High School. A former colleague of mine reports having been *very* bored in high school, and admits to frequently skipping classes ... so frequently, in fact, that she ought to have been kicked out of school. She never *asked* for an exception to the rule that gets you expelled for skipping class, but an exception was made: one to which she owes her college education, and so her current job, and so, indirectly, much of the shape of her current life. In this example, the effect of not having an exception made for one is rather more dramatic than in the first.

The Tardy Contributor. A fellow academic who was editing a *Festschrift* had set a hard deadline for the invited papers. One of the authors circulated a draft of his paper, and shortly before the deadline discovered a problem requiring major revisions. The editor granted an exception to the deadline; without it, the author would have had to withdraw the paper, which would have hurt the *Festschrift*, the feelings of its subject, and the author himself, who would have been hard put to find another venue for the commissioned piece.

Cases like these are recognizable enough (although we still need to take up the question of how they are to be interpreted). But are they common enough to make the availability of exceptions a precondition for successful agency? If the need for exceptions is only exceptional, it will not support the argument we are developing against universalizing the CI-procedure. So I will now argue that exceptions will be needed on an ongoing basis.

Let's begin with a fact used by other Kantian arguments that we've already reviewed: that human agency is dependent on the cooperation of others. Human projects are vulnerable, the kind of projects that human rational deliberators reasonably adopt will outrun the resources that an agent can muster on his own, and they will do so frequently enough to make cooperation a *sine qua non.* Once again, this is not being introduced as a necessary truth about agency. For all we know, there could be agents who were successful lone wolves, either because projects that wolves take on fall into a narrower range than ours do, or because wolves are much more resourceful than we are.

The behavior of others is a very large part of the environment in which we pursue our ends, and much of that behavior is policy-directed or rule-governed in any case; ex hypothesi, in the world of the universalized CI-maxim, *all* of it is. So the contribution that others make to our projects will (mostly, in the actual world, and entirely, in the perturbed social world we are considering) come under the heading of rules or policies in force. So you will be able to pursue your projects effectively (without getting exceptions to the rules) only if policy- and rule-governed behavior gives you the cooperation you need. So we need to ask how likely that is.

It is obvious that your projects have to be chosen largely in ignorance of others' rules and policies. For one thing, most people's policies are unannounced. And perhaps more importantly, even if they were announced, no one could keep track of more than a handful of them. This is a special case of another empirical claim used in the Kantian argument for developing one's talents: that people can't develop all the capacities or skills they will need. The skill in this case is that of knowing what the rules of the game are; lawyers are a class of professionals who specialize in developing that skill for a smallish subset of the official rules of the game, and becoming competent in just such a small subset turns out to be a full-time occupation. In special cases, you may consult a specialist—a lawyer or accountant—before embarking on a project, but most of the time, that's just not feasible. Because it is typical of interesting or important projects that one doesn't know just how they will unfold, one doesn't even know whose rules one will run into along the way. (This is true of not-so-interesting projects as well; in the parking example, I did not know, when I made my travel plans, that I would have a problem leaving the car on the street.) As an agent, you choose your direction with only the most sketchy sense of what the other relevant agents' policies are likely to be.

It is as obvious that both individuals and institutions have to formulate their policies and rules in ignorance, for the most part, of *your* ends. It is not just that it wouldn't be logistically feasible to keep track of everyone's ends. In the Kantian picture of rational agency, the central feature of such agency is that you can formulate and adopt new ends. You yourself can't predict what your own ends will be down the road (recall that the New Kantian arguments for mutual aid and for developing one's talents rely on this fact); a fortiori, others cannot predict what your ends will be either, when they are considering what rules to adopt.

Since each set of decisions—about what rules and policies others adopt, and about what ends and projects you adopt—is made in ignorance of the other, the chances of their being suitably coordinated are very small indeed. Overall, we should not expect that others' policies will, as they are, deliver the cooperation that your projects (and you) need from them. Therefore, a condition on the successful exercise of your own agency is that others make exceptions for you. It is not for nothing that unions use work-to-rule as a threat.[26]

Let me pause to address a handful of worries. First, you may be worrying that parking problems and their ilk are too small to drive a criticism of Kantian morality. I certainly agree that we should not reject Kantian moral theory because on some occasion I could not find a place to park, but the smallness of such examples is meant to serve as an icon for what the world of human agency is really like. The pervasiveness of such small problems means that, although one can probably do without an exception in this or that particular case, doing without exceptions in all of them will make one's remaining agency not worthy of the name.[27]

Second, you may also be worrying that the examples are not *moral*, or that the exceptions in my examples are undeserved. But the Kantian argument (on its New Kantian reading) is supposed to be driven by one's own stake in one's agency, and so that is what matters for the argument, not whether the subject matter strikes one as moral, and not what one does or does not deserve. Objections turning on desert and on what is and is not moral get the order of explanation backwards, because one of the great strengths of Kantian theory is that it purports to provide a criterion for inclusion in the subject matter of morality: to appeal to one's independent and prior view of what is a moral issue and what morality requires is to beg at least one of Kant's questions.

Third, the argument I have been developing, like other arguments for imperfect duties, requires one to make judgment calls about *how often* a given type of situation arises. My own judgment call is that the relevant situations really do arise often enough for your agency to depend on exceptions granted by others (and, although I have not emphasized this side of the argument, exceptions to your own policies granted by yourself—think of a landlord granting himself an exception to a self-imposed rule that prohibits him from renting to dog owners). But if you have not been convinced, recall that the argument is modeled on the Kantian arguments for mutual aid and for the development of one's talents; it will suffice for present purposes if one needs

exceptions from others to roughly the same extent that one needs assistance from others. This is much harder to gainsay: after all, one can go days or weeks at a time without needing to ask for anyone's help. That is, the fallback claim is that if the model New Kantian contradiction-in-the-will arguments work, so does this one.

In any case, this line of resistance has an analogue discussed by Kant, who remarks that a well-off individual is likely to be happy to forgo the promise of mutual aid.[28] I have noticed that those who think we do not need exceptions as often as all that tend to be the high-SES academics; that is, there is a recognizable class bias to the objection. What is more, even observations that cut across class lines may understate the overall need for exceptions, because some regions (think First World economies) accumulate exceptions at the expense of faraway and less-developed parts of the world. (If the global ecosystem can't handle an SUV for everyone, and if Americans predominantly drive SUVs, then Americans are collectively taking an exception.) In particular, economic surplus that can make exceptions seem unnecessary is itself a giant, economy-size exception. I expect that the insistence that exceptions are unnecessary often marks a deep sense of entitlement that accompanies their being consumed unnoticed.

Let me field one further worry before moving on. If the perturbed social world of the universalized CI-maxim is one in which people act in accordance with the Categorical Imperative, then they will not lie, will lend a hand more frequently and with greater alacrity than is actually their wont, will have skills that make their assistance more effective than it now is, and so on. That world will be a much kinder and gentler place than ours, and even if exceptions are necessary for agency in our world, perhaps they will not be required in what we might as well call the Kingdom of Ends. But if they are not, then the contradiction in the will we have been pursuing is avoided.

Here we are really being asked to choose between two versions of the CI-maxim, one of which has us take such effects into account (to produce what we can call a highly perturbed social world), and the other of which damps out the effects of the universalized maxim much more quickly (call its product the minimally perturbed social world). Recall that the point of the CI-maxim (expressed in its final clause) is to determine whether some line of action that you are actually contemplating is permissible. But the highly perturbed social world is too distant from the actual world to allow you to assess the actions you must in fact consider. Suppose that the maxim you are contemplating is: to fill up your

tank when you pass a gas station, so as not to run out of gas. Cars predictably kill and maim some non-negligible percentage of their passengers, so in a Kantian choice of transportation system, automobiles would be prohibited; in the Kingdom of Ends, there are no cars. So you cannot turn to the Kingdom of Ends to answer questions about when to tank up. Appealing to the highly perturbing version of the CI-maxim to determine what to do is adopting what is normally a self-frustrating plan of action: it amounts to either a contradiction in conception or something on a par with one.

3.

Until this point in the argument, New Kantians will find it easy to be concessive. It is hard not to allow that there are many cases where one needs what the man in the street calls an exception to the rule. But Kantians will have a complaint to register about the interpretation I have been putting on them: that they are exceptions in one sense, but not in another (and in particular, not in the sense my argument requires). Korsgaard has put the response in print, and so I will use her as my stalking horse; however, it is important to bear in mind that I am taking up her discussion not just as an objection one could make to my argument, but as an instance of an objection New Kantians *have* to make: this is a forced move.

Korsgaard agrees that

> 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 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.*[29]

The way she tries to accommodate this very pervasive fact is to allow that one's principles "be willed ... as provisionally universal," which is to say that we are to "think [a principle] applies to every case of a certain sort, unless there is some good reason why not." She invokes the Kantian comparison to causal laws, and reminds us that when we make causal claims, we usually invoke causal principles that hold only "all else equal." When we find an exception to a natural law, "we look for an explanation. Something must have made this case different: one of its background conditions was not met."

Here the important point is that the explanation must itself have universal (or provisionally universal) force. For this reason, exceptions can be incorporated into laws that are universal in form; and this

requirement (or the analogous requirement, since maxims really do differ from natural laws) in the practical case runs as follows:

> if a principle was provisionally universal, and we encounter an exceptional case, we must now go back and revise it, bringing it a little closer to the absolute universality to which provisional universality essentially aspires.

That is (and adjusting the terminology to bring it into line with the discussion so far), while (New) Kantians can allow that you may encounter cases that, you will agree, aren't properly handled by whatever universally shaped maxim you had adopted, what you are to do is not, strictly speaking, to grant an exception, but to replace your old maxim with a new one that handles the "exception," as it really ought to have been treated by the rule in the first place.[30]

This might be done by tacking the exception onto the maxim, as an extra clause, so to speak (one that you might not bother to mention the next time you state your maxim, but that is now understood to be part of it). Or you might reformulate the rule as a whole, so that separate mention of the class of exceptions wouldn't be needed. There are many ways, some more and some less elegant, by which this might be accomplished.[31] What they share is an adjustment in the contours of your practical commitments, and what matters is that while this adjustment may be very sensitive to, and picky about, details, it too is universal in form: you are now committed to handling other similar cases in the same way.

Now Korsgaard thinks that "[t]he difference between regarding a principle as universal, and regarding it as provisionally universal, is marginal." This, I am going to argue, is a mistake; the difference makes all the difference in the world, and especially in the world of the universalized CI-maxim. There are two things to notice here: first, that making an exception and building an exception-shaped twiddle into a rule work differently in the social world; and second, that the commitment to revise one's maxim in such cases is itself a maxim, and has to be checked against the CI-procedure as well.

4.

Recall the Kantian argument against lying: if lying in given circumstances was a practice, everyone would know that it was, and they would adjust their behavior in the light of that knowledge. (That is, no one would believe your lies, and you would never get that loan.) More gen-

erally, when working our way through the CI-procedure, we have to assume that others are going to adjust their plans of action in light of what the rules are. Recall that I left open the question of whether this was to be regarded as an empirical fact, to be retained in the world of a universalized maxim, or a prescriptive constraint—a matter of how Kantian moral theory requires us to think about our fellow persons. I do not want to resolve this question now, so I will try to use cases for which both versions of the constraint are plausible.

It follows that there is a difference between a genuine exception, and an adjustment to a rule that accommodates a previously unnoticed class of cases: when thinking about the latter, Kantian theory requires us to consider how people will adjust their plans in light of the modified rule. (People who grant exceptions sometimes worry about setting precedents; another way to characterize the effect that now has to be taken into account is that Kantian exceptions *always* set precedents.) Let's return to our previous cases:

Parking in Milan Revisited: If everyone who needed to were allowed to leave his car in the lot with his blinkers on, people would soon notice this option and start planning around it. They would count on being able to dash upstairs to make a phone call, or to deliver flowers, or whatever; and so they would plan on making that phone call or delivering the flowers. The parking lot would very quickly be full of briefly parked cars. The primary users of the lot would not be able to get in and out, and the visitor who needed an exception to the rule would generally find all the free spaces already taken. That is, what I needed to make my day work was an *exception*, and not an adjusted rule.

Trouble in High School Revisited: If it became the rule, roughly, that promising and bright high school students got to skip class with impunity, they would soon realize that. An important incentive for attending classes would disappear, better students would stop coming to class, less promising students would be transformed into resentful second-class citizens, and the effect would be, not to help out bright and promising students (the intent of the original exception), but to undermine their academic performance. Again, what is needed is an exception, rather than a modification to the rule.

The Tardy Contributor Revisited: Likewise, if it were discovered that deadlines could be broken, contributors would quickly realize this. And in fact they have; some authors routinely overbook themselves, knowing that deadlines do not have to be taken seriously. (If you're in the business, I don't need to tell you who the egregious abusers are.)

Any volume with a sizable number of contributors is bound to have one or more such authors, and since other authors know that the volume will be delayed, they too plan on dawdling. Since no one really believes the volume will have a timely appearance, authors reserve their best efforts for other venues. The point of granting the original exception was to facilitate the appearance of a strong and timely collection, which point is precisely undermined by universalizing the exception. Again, what is needed is an exception, rather than a modification to the rule.

More generally, because rational agents adjust their plans to take account of changes in the rules (because they, quite correctly, come to *count* on the new rules), institutionalizing exceptions—writing them into the rules in the way that Kantian moral theory requires—tends to have perverse results. The results are perverse in the ordinary sense, that of giving rise to baroque and unwanted side effects (think of the tax code, to take a bureaucratic example); but they are also perverse in a technical and Kantian sense, that of undercutting the connection between the exception and its originally intended effect. Usually, granting an exception while institutionalizing exceptions of that kind amounts to what is, by Kantian lights, a self-frustrating plan of action.

There are of course two sorts of cases: those in which universalizing an exception undermines the point of granting such an exception, and those in which it does not. New Kantian arguments for imperfect duties turn on judgment calls (about how much of an impediment to agency a given social phenomenon will turn out to be); my sense of the territory here is that you can't get by just on the latter sort of exceptions. But once again, if your sense of the territory differs, the backup claim is that such undermining happens to roughly the extent that you turn out to need help (and skilled help) from other people—that is, that if the New Kantian model arguments work, so does this one.

5.

Our first objection to treating maxims as "provisionally universal" was that there is a substantive difference between exceptions and changes in the rules, and often agents need the former rather than the latter. Our second objection is located on the side of the person dispensing exceptions, rather than on the side of the recipient. Kantian strategies need to be subjected to the Kantian consistency test. Accordingly, consider the following second-order maxim:

> When I run into a case that my (first-order) maxim does not handle properly, I will revise my (first-order) maxim to incorporate the exception, so as to be able to act on rules that I understand to be "provisionally universal."

Call this the *Revision Maxim*. If a Kantian is to find the Revision Maxim acceptable, it must pass the CI-procedure; but there is a quick argument to show that it does not.

I have for expository convenience been speaking of agents adjusting themselves to changes in the rules, and recent Kantians may have found that turn of phrase worrisome; what we are supposed to be imagining is not a response to change, but the steady-state social world of which the contrary-to-fact alteration is a permanent feature. (See, for example, Rawls 2000, 171f.) Now in the steady-state world where agents follow the Revision Maxim, most of their maxims will *already* be much-revised. That is to say, they will be very complicated (in something like the way the legal code or the tax code is complicated). But it is a deep empirical fact about human beings—a fact on a par with their vulnerability and the limitations on how many useful skills they can acquire—that they are not very good, cognitively, at handling this kind of complexity and detail. It is very easy to swamp human cognitive resources, and in a world in which one's maxims have had many exceptions appended to or integrated into them, they will be too complex to think with.

First, it will not be realistic to demand of agents that they be sensitive to whether their maxims are universalizable, because that will usually be too complicated a question for them to answer. The perturbed social world of the CI-procedure is a complex abstract object, and you can think of it as being a little like a chessboard. Just as it is very easy to produce boards that are hard to see your way through (the starting state of the board is hard for anyone, including grandmasters and IBM's Deep Blue; that's why people *play* chess), so it is easy to produce perturbed social worlds it's hard to see your way through. Adding detail to rules makes it harder to see, and see one's way through, the structure of the perturbed social worlds they generate. (Compare the way the tax code evolves: policy makers fail to see that a provision they are introducing will create new loopholes, and that these will be exploited to the hilt; they then need to add further provisions, which in turn create further loopholes; and so on. If human beings could generally see what the perturbed social world of a relatively complex rule looked like, they

would be able to do so when the subject was taxation, and the tax code would not have the daunting patch-on-top-of-patch look that it does.) The point of the Revision Maxim is to allow one to act on rules that are universalizable; but in the world of the universalized Revision Maxim, one often cannot tell whether the maxim one proposes to act on is universalizable.[32]

Second, as maxims get more complicated, it becomes harder to see what actions they demand. (Think again of the tax code; who among us, accountants included, knows what taxes we should really be paying? If human beings could see what very complicated rules require, they could do it when the subject was taxes.) Again, the point of the Revision Maxim is to allow one to act on rules that are universalizable; but in the world of the universalized Revision Maxim, one often cannot so much as tell what one's rule requires that one do. If this is right, the Revision Maxim fails the CI-procedure (twice over) at the contradiction-in-conception stage.[33]

Kant himself seems not to have appreciated the problem here; he apparently thought that making out the demands of the Categorical Imperative, first formulation, was within the reach of even the simplest intelligence. Kant has not been alone. Widespread awareness of complexity as an obstacle to problem solving dates only to the 1960s, and even now the idea has just barely become respectable within economics and political science.[34] For instance, with occasional exceptions such as Hayek, the insuperable difficulties of managing centrally planned economies were overlooked until very late in the twentieth century (and it is suggestive that the construction of a Soviet-style Five Year Plan bears a family resemblance to the mapping out of perturbed social worlds that we are now contemplating).[35] If the realization that it is often not cognitively possible to solve problems about complicated social structures is still being assimilated in the social sciences, we should not blame Kant for not having noticed it.

You may be inclined to think that the complexity of the perturbed social worlds generated by the CI-procedure cannot be as intractable as all that. After all, there are domains—such as producing syntactically correct speech—that seem to be quite complex and in which humans do quite well.[36] And you may be wondering whether perhaps the CI-procedure belongs in such a domain. To see why this is unlikely, contrast the almost effortless production of sentence after grammatical sentence (that gives the Chomskian hypothesis of a special-purpose hardwired grammar module whatever plausibility it has) with the pau-

city of worked examples of the CI-procedure. Earlier on, I added the New Kantian argument against violence to the three examples that canonically accompany presentations of the Categorical Imperative. Despite the fact that violence is a very basic moral issue, it took Kantians some two hundred years to come up with the argument, and it is treated, in the New Kantian literature, as a noteworthy achievement; it is, in any case, one of the very few new applications of the CI-procedure. If humans came to the problem space generated by the CI-procedure possessing anything like their innate competence with the syntax of natural languages, then the solved problems would not be nearly so few and far between, and this particular application would have been merely one more of a practically endless stream of them.

Even if treating maxims as provisionally universal is unworkable, it might still be tempting to think that the pressure for exceptions can be relieved, in a way Kantians would welcome, by building generic exception clauses of one kind or another into the contents of maxims themselves. (For instance, "... unless an emergency comes up," or "... other things being equal," or "... unless an exception is needed.") An argument of the kind we have just assembled disposes of these proposals. When it comes to "other things being equal," there are indefinitely many such cases that might come up, we cannot estimate their frequency, and they do not lend themselves to being neatly classified in advance. So when maxims contain such clauses, it is impossible to tell what their perturbed social worlds will look like. For maxims that have had their contents blurred in this kind of way, the CI-procedure does not give definite results. The point of adopting a second-order maxim that dictates generic exception clauses in one's first-order maxims would be to rescue the CI-procedure. So such a maxim founders on a contradiction in conception. This is why agents' maxims have to be treated as laying down exceptionless policies.

To recap: First, modified rules differ from genuine exceptions, and they tend to be self-frustrating when universalized. Second, acting on the Revision Maxim is what New Kantian moral theory requires of agents who make exceptions, and the Revision Maxim fails the CI-procedure. This means that New Kantian moral theory cannot accommodate the need for exceptions. If Kantians cannot after all make exceptions, and if, as I have argued, exceptions are a necessary precondition for successful agency worth the name, then the CI-maxim does indeed give rise to a contradiction in the will.

6.

Before concluding the paper, I want to take up two related objections to the argument we now have on the table. These try to abort it by circumscribing the application of the concept of a maxim. To frame the objections, I am going to take a moment to introduce another element of the New Kantian picture, which I will call *maxim hierarchies.*

The third clause of a maxim, as the New Kantians construe it, specifies the end or point of one's action: to revert to an earlier example, "when I pass a gas station, to top off my tank, *so as not to run out of gas.*" Now, that point can in general be understood as a further action or plan of action; the intent of that further plan of action can in turn be rendered by a maxim; and that further maxim will itself specify a further point: perhaps, "when I am driving a car, not to let myself run out of gas, *in order to keep my car a reliable means of transportation.*" And likewise, that further point can be unfolded into a still-further maxim: maybe, "when I am in a location that does not have convenient and dependable public transit, to keep my car a reliable means of transportation, *so that I can get where I need to go, when I need to be there.*" Maxims that appear at the upper reaches of such hierarchies tend to specify more abstract and more general policies (and I'll sometimes talk about them as dictating "larger" actions, because they contain the actions specified by their inferiors in their hierarchies as subplans or components).[37] Such hierarchies turn up in the exposition of most New Kantian positions,[38] and they give a claim I made earlier the status of an observation: what counts, for present purposes, as action does not have to look *busy*; in particular, suitably governed omissions or refrainings can count as actions. In order to keep my car a reliable means of transportation, I don't strip the gears. (And so I don't shift directly from fifth to first; I don't shift into reverse while I'm speeding down the highway; and so on.) Not stripping the gears occupies a place in (another branch of) my maxim hierarchy, even though it, and many of the actions below it, are things I *don't* do. (Recall, under this heading, that testing a general policy of *not* helping was the Kantian way of arguing for a duty of mutual aid.) For the purposes of practical reasoning, deliberatively governed plans of inaction are plans of action, too.

Now we can state the objections. First, at the upper reaches of such a hierarchy of intentions, there is a glass ceiling above which the policies are not to be considered maxims. In particular, being guided by the CI-procedure need not be a maxim, and so such a policy does not

itself need to be tested using the CI-procedure. But if we exempt what we were calling the CI-maxim (and similarly general or abstract policies, such as the Revision Maxim) from passing the CI-procedure, then the argument we have just finished constructing will not go through. Since maxims were introduced as generic representations of the contents of intentions or volitions, this amounts to saying that, while you are to act in conformity with the dictates of the CI-procedure, this is not to be taken for a further intention on your part: in an older Kantian locution, you are to act *in accord with*, but not *from*, the moral law. Call this the *Upper Glass Ceiling Objection.*

The Upper Glass Ceiling Objection does not sound much like Kant, who *is* very concerned that one act, not merely in accord with, but out of respect for the moral law; Kant explicitly characterizes as a maxim an extremely general and abstract policy, "that I should follow such a law even if all my inclinations are thereby thwarted" (G 400); he also worries about how maxim hierarchies are going to top out, treating the Categorical Imperative as one of the possible basic postures an agent might assume.[39] So the Upper Glass Ceiling Objection would have to be understood not as an explication of Kant, but as an amendment to him. Now, the CI-procedure is motivated by the idea that it is important to assess, not your actions on their own, but why you do them; if you did not think that what mattered was the practical consistency of the volition, rather than the outcome effected by the volition, there would be no point to deploying the CI-procedure in the first place. So it's hard to see why someone who cared only about conformity of reasons and rationality to some template would end up requiring *this* pattern to be conformed to in one's mode of producing action. That is, the move made by the Upper Glass Ceiling Objection is evidently motivated not by considerations that have their home in the (New) Kantian way of seeing things, but as an ad hoc response to the problem posed by our argument. Of course, that is not yet reason enough to dismiss it.

Second, it will be suggested that what we find at the lower reaches of the hierarchy are not, properly speaking, maxims. Only suitably general principles, such as those mandating or prohibiting lying, are subject to test by the CI-procedure; "specific intentions" are not,[40] and so need not be exceptionless. Since the argument we have been constructing turns on the pervasiveness of needed but unavailable exceptions, if a glass ceiling below which maxims are not to be found is set suitably high, then the argument will not go through. Call this the *Lower Glass Ceiling Objection.*

The Lower Glass Ceiling Objection requires a Kantian willing to make a difficult sacrifice. Hegel complained that Kantian universalizability was too formal a device actually to produce substantive results, and the New Kantian tradition takes visible pride in having shown how a recognizably Kantian rendering of the Categorical Imperative can indeed produce conclusions as concrete as those we have reviewed. But the further up the Lower Glass Ceiling is set, the fewer substantive results there will be. Moreover, this New Kantian rejoinder puts a great deal of pressure on the notion of a maxim. If maxims are not simply a way to capture the form of any intentional action whatsoever, we need to be told what they are. And looking more closely at this lacuna gives us a decisive response to both the Upper and Lower Glass Ceiling Objections.

Recall the structure of Kantian imperfect duties. No one can help everyone who needs it, and there is no way of marking off just those cases in which you really have to help out from those in which you don't. So it is up to you to decide when to help out; in the end, when to help is your judgment call. Now, it is not as though we have a principled way to place what I was calling the Upper and Lower Glass Ceilings; after all, what is the hard and fast difference between maxims, more specific intentions, and more general policies? The CI-procedure applies only to what counts as a maxim, so the Glass Ceiling Objections give the moral law the structure of an imperfect duty: they entail granting to agents the discretion to decide when to invoke the CI-procedure as a constraint on their actions. (So an agent might insist, entirely legitimately, that lying was permissible because, in his judgment, the intention to lie was either too specific or too general to count as a maxim.) The effect would be to erase the distinction between perfect and imperfect duties from New Kantian moral theory. I am confident that the New Kantians would regard that consequence as unwelcome and unacceptable.[41]

New Kantians tend to allow that maxims can be very general and abstract.[42] But some New Kantians already think that intentions at the lower regions of the hierarchy cannot in any case count as maxims. So is my objection to Lower Glass Ceilings uncharitable?[43] In my view, New Kantians face a hard choice, and here is how I think it looks.

When running a maxim through the CI-procedure produces a startling result, there is an unfortunate inclination on the part of Kantian theorists to try to block the result by insisting that the input was not really a maxim in the first place.[44] I am for my own part disappointed

by how frequently that inclination wins out. Kantian theory is deep and interesting only when it is taken seriously, and I have found that facing up to the startling result is usually theoretically fruitful. Just for instance, the maxim, "I will buy clockwork trains but not sell them," does not universalize: if no one sells the trains, no one will able to buy them (Nell 1975, 76). The moral really is that *that* intention is impermissible. (Morality requires more carefully conditionalized maxims, perhaps ones that are explicitly sensitive to the idea that being a toy collector is only one of the many social roles you might contingently occupy.) Or again, it's a standard classroom example that maxims like "I will turn up at Times Square tomorrow" don't universalize. (There's not enough room for everyone.) The moral here is that introducing proper names into maxims produces contradictions in conception. Or again, Herman's example of a "puzzle maxim" that must be "set aside" is, "To always be first through the door." I disagree: the fact that this maxim is not universalizable is a very good candidate explanation for the fact that this sort of pushiness is, while not a big deal, nonetheless *rude.* That is, one option for the New Kantians (and the one that seems to me likeliest to produce richer results) is not to give up so easily on the intentions at the bottom of the hierarchy (but this means giving up on the Lower Glass Ceiling Objection).

A second option is to find a principled way of distinguishing the maxims to which the CI-procedure applies from the more concrete intentions to which it does not. This seems less promising to me. One reason is that no one who has tried to introduce the distinction has made much headway at cleanly articulating it. Another is that the New Kantian tradition has seen the CI-procedure as a test of practical consistency. That way of motivating the CI-procedure makes sense when its range of application is intention, generically understood. But what kind of distinction could support a more narrow notion of consistency? And why should practical consistency be required only of some special type of intention?[45]

The third option is to allow perfect duties to vanish from New Kantian moral theory, and the centrality to the tradition of the perfect duties makes this tantamount to altering the content of Kantian moral theory almost beyond recognition. But if New Kantians cannot afford glass ceilings, then they are committed to maxims coming in all degrees of generality and abstraction. The New Kantian position may require that maxims share a shape—the logical form that is used to express the content of an intention that comes up to snuff—but their

shape cannot constrain the level of abstractness of the maxim's content. And this is (one more reason) why I have been using the looser 'point' to describe what the third slot of New Kantian maxims expresses; some people are finicky about the term 'end', and prefer to use it only for fairly concrete aims with definite termination points. But this notion of end is much narrower than what maxims must be able to capture.

<center>7.</center>

What lesson should we draw from the argument that we have just concluded? Not the easy one, that we should give up on New Kantian moral theory. True, if contradiction-in-the-will arguments establish imperfect duties, it looks as though we have exhibited a Kantian imperfect duty to *violate* the Categorical Imperative, and so that part of the position will have to go. But New Kantian moral theory would not have gotten nearly the attention it has if it were so intellectually impoverished as to be a one-idea view. I mentioned at the outset two other ideas traditionally identified with Kant (the requirement that persons be treated as "ends in themselves," and the importance of autonomy), but there are many more: just for instance, the conception of personhood as a practical rather than a metaphysical status, ingenious arguments against instrumentalist accounts of practical reasoning, and the suggestion that actions are to be thought of as moves in the only game in town. (See Korsgaard 1996a, chaps. 13, 11, Schapiro 2001.) Showing that one of the ideas in the Kantian portfolio is unworkable leaves a valuable and still-diversified grouping of philosophical assets. (And in fact some New Kantians have over the past decades come to rest more weight on the so-called Formula of Humanity in particular.)

It is also too early to go looking for constructive lessons about practical reasoning and morality—the kind of lessons that would help us frame improved accounts of one or the other. We ought first to develop a deeper diagnosis of what has gone wrong with the Categorical Imperative, first formulation; but to come by that, we will need a better handle on the philosophical motivations of the New Kantian position than we now have, starting with an explanation of why an agent wills the universalization of his maxim (so that the contradictions exhibited by the CI-procedure are contradictions *in* the agent's will). I do not think this question can be successfully pursued without turning

from the New Kantians back to Kant himself, and that is an undertaking for another time and place.

So allow me to suggest an interim and methodological lesson. When you are working up your philosophical theory, always stop to check what happens when you apply that theory to itself. Sometimes that operation will not so much as make sense. Sometimes the result will be fast and reassuring. But sometimes, as we have just seen, it will not, and so the test of reflexive application is not one that you can afford to neglect.

University of Utah

References

Anderson, R. Lanier. 2001. Synthesis, Cognitive Normativity, and the Meaning of Kant's Question, 'How Are Synthetic Cognitions A Priori Possible?' *European Journal of Philosophy* 9:275–305.

Anscombe, G. E. M. 1985. *Intention.* 2nd ed. Ithaca: Cornell University Press.

Bendor, Jonathan. 2003. Herbert A. Simon: Political Scientist. *Annual Review of Political Science* 6:433–71.

Bok, Hilary. 1998. *Freedom and Responsibility.* Princeton: Princeton University Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press, Clarendon Press.

Hayek, Friedrich. 1948. *Individualism and Economic Order.* Chicago: University of Chicago Press.

Hayek, Friedrich. 1989. *Order—With or Without Design?* Ed. Naomi Moldofsky. London: Centre for Research into Communist Economies.

Herman, Barbara. 1990. *Morality as Rationality: A Study in Kant's Ethics.* New York: Garland. Reprint of 1976 doctoral dissertation.

———. 1993. *The Practice of Moral Judgment.* Cambridge: Harvard University Press.

Hill, Thomas. 1992. Making Exceptions without Abandoning the Principle: or How a Kantian Might Think About Terrorism. In *Dignity and Practical Reason in Kant's Moral Theory,* 196–225. Ithaca: Cornell University Press.

Kant, Immanuel. 1781/1787/1998. *Critique of Pure Reason.* Trans. Paul Guyer and Allen Wood. Cambridge: Cambridge University Press.

———. 1785/1981. *Grounding for the Metaphysics of Morals.* Trans. James Ellington. Indianapolis: Hackett.

———. 1788/1997. *Critique of Practical Reason.* Trans. Mary Gregor. Cambridge: Cambridge University Press.

————. 1793/1998. *Religion within the Bounds of Mere Reason and other Writings.* Trans. Allen Wood and George di Giovanni. Cambridge: Cambridge University Press.

————. 1797/1994. Metaphysical Principles of Virtue. In *Ethical Philosophy,* 31–161. Trans. James Ellington. Indianapolis: Hackett.

————. 1902–. *Kants gesammelte Schriften.* 29 vols. Ed. German Academy of Sciences (formerly the Royal Prussian Academy of Sciences). Berlin: de Gruyter.

Korsgaard, Christine M. 1990. *The Standpoint of Practical Reason.* New York: Garland. Reprint of 1981 doctoral dissertation.

————. 1996a. *Creating the Kingdom of Ends.* New York: Cambridge University Press.

————. 1996b. *The Sources of Normativity.* Cambridge: Cambridge University Press.

————. 1999. Self-Constitution in the Ethics of Plato and Kant. *Journal of Ethics* 3:1–29.

Nell, Onora. 1975. *Acting on Principle: An Essay on Kantian Ethics.* New York: Columbia University Press.

O'Neill, Onora. 1989. *Constructions of Reason.* Cambridge: Cambridge University Press.

Rawls, John. 1989. Themes in Kant's Moral Philosophy. In *Kant's Transcendental Deductions,* ed. Eckart Förster, 81–113. Stanford: Stanford University Press.

————. 2000. *Lectures on the History of Moral Philosophy.* Ed. Barbara Herman. Cambridge: Harvard University Press.

Schapiro, Tamar. 2001. Three Conceptions of Action in Moral Theory. *Noûs* 35:93–117.

Trudeau, G. B. 1996. *Virtual Doonesbury.* Kansas City: Andrews and McMeel.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59:433–60.

Vogler, Candace. 1995. Philosophical Feminism, Feminist Philosophy. *Philosophical Topics* 23:295–319.

————. 2003. *Reasonably Vicious.* Cambridge: Harvard University Press.

Notes

[1] Rawls 1989 is an overview that appropriates the position it attributes to Kant as a precursor of Rawls's own constructivist political theory; Rawls 2000 makes available his very influential lectures. It should not be assumed that the influence was all one-way; over the years, his presentation of Kant assimilated much of the work he had delegated to his students.

Nell 1975 focuses on laying out the Kantian deliberative procedure, providing a much less terse rendering of the view set out in O'Neill 1989, chap. 5 (and more generally throughout the volume). Korsgaard 1990 accepts O'Neill's account of the procedure and focuses on motivating it; Korsgaard 1996a develops a range of Kantian positions around the motivated procedure, and is now probably the center of gravity of this reading. While Herman 1993 advances the interpretation, it also breaks ranks on a number of points, and in section 6 I will discuss the reasons Herman's views diverge from others in the group.

I will not tie myself too tightly to the nuances of any version of the New Kantian view; I mean my sketch of it to represent the shared structure fairly, but I do not want to imply that the fine print is attributable to each and every instance of the class. I will flag important disagreements as I go, as well as descriptions of their view that New Kantians would themselves find controversial or prejudicial.

[2] From G 421, with grammatical modifications. I will cite Kant's works using the following abbreviations, in the translations following the title: G: *Grounding for the Metaphysics of Morals* (1785/1981); A/B: *Critique of Pure Reason*, giving the pagination for the A and B editions (1781/1787/1998); C2: *Critique of Practical Reason* (1788/1997); DV: *Doctrine of Virtue* (*Metaphysics of Morals*, second part) (1797/1994); R: *Religion within the Boundaries of Mere Reason* (1793/1998).With the exception of the first *Critique*, pagination will be that of the standard Academy edition of Kant's works (1902–).

Herman 1990, 150f., 188, provides a useful compilation of the different formulations of the Categorical Imperative.

[3] Despite the step-by-step presentation and occasional remarks by the authors we are considering (Herman 1993, 115, calls it an "algorithm"), the CI-procedure is not technically a procedure or algorithm. Procedures can be executed mechanically and are guaranteed to terminate; inspecting the steps in this one will show that neither is true of it; Rawls 2000, 166, acknowledges this point, as does Nell 1975, 73. Rawls's own work progressively distanced itself from the idea that the central problems of political theory could be rendered as well-defined exercises in game theory, and I suspect that the "procedure" terminology is a holdover from an earlier stage in his development.

[4] Rawls adds a fourth "unless" slot that we can understand to be included in

C (2000, 168). Kant and the majority of the New Kantians think of the point of the action as its end; Kant in particular holds it to be a formal fact about actions that they have identifiable ends (C2 34, DV 382–85). I am being looser about this partly to accommodate Herman (1993, 221), who quite plausibly insists that the full desirability characterization of an action is to be represented in its maxim, and partly for reasons I will get to in due course.

[5] I'm framing the condition disjunctively because it's not always clear or agreed among the New Kantians which disjunct the test pivots on. See Nell 1975, 69ff;. Herman 1993, 118; Korsgaard 1996a, chap. 3, esp. sect. 3; Rawls 2000, 169; O'Neill 1989, 96.

[6] Herman is something of an exception; see note 45, below.

[7] G 421 n., DV 390, 392–94. I don't here want to take up the question of how plausibly the perfect/imperfect distinction can be tied to the step at which the CI-procedure is exited. For a New Kantian review of exegetical controversies having to do with this distinction, see Nell 1975, chap. 4.

[8] The point now is not that a practice that is unsuccessful for this kind of reason will be replaced by a different practice. As a matter of fact, we more or less live in the world Kant describes, where people lie to get loans as a matter of course. That is why loans are not made on the basis of such promises. Mortgage originators do not stay in business by trusting their customers; they make sure there is collateral to foreclose on.

[9] Closer to home, letters of recommendation for the academic job market are another case where, for the most part, the discounting doesn't go nearly deep enough. "His dissertation will certainly change the field he is in." "She is the best student who has ever come through our program." "The work he has done as a graduate student would make a successful case for tenure." Sound familiar?

[10] Because Kant predates the twentieth-century repudiation of psychologism, claims with prescriptive force are often presented in his work as descriptions of an idealized mind. For some discussion, see Anderson 2001.

Presentations of the CI-procedure usually include a gloss to roughly the effect that in constructing your representation of the perturbed social world you are to keep as much as possible of the world as we know it intact. This is analogous to the way we reason about contrary-to-fact conditionals; when asking "What would have happened if *p*?" we imagine as much as possible of the background to remain as is. We have just seen that it is still an open question whether this is the appropriate understanding; however, see also the final objection in section 2, below.

[11] Is it *really* conceivable? Perhaps not: a feminist objection to Hobbesian state-of-nature arguments is that human beings are too vulnerable actually to grow up or live in such a world (Vogler 1995).

[12] Again, this could be taken as an empirical claim, or as prescriptive, about how your choice has to look "from the practical point of view." (For development of this latter idea, see Bok 1998.) The prescription can be motivated by a specifically moral thought such as: just as you have to respect others, by leaving them space to adopt a reasonable range of ends, so you have to respect yourself, and not allow that range of ends to be foreclosed to you. The price of

this kind of motivation, however, would be building explicitly moral consider-
ations into the conception of practical rationality at the core of the theory.
(I'm grateful to Geoff Sayre-McCord for discussion here.)

[13] Again, is it really? It is suggestive that, while Kant seems to have thought
that the no-talent maxim guided the way life was lived in the South Pacific, no
actual human society has ever operated on this basis.

[14] This particular argument may not seem much like Kant's (somewhat
obscure) text, but it is a New Kantian way of handling the case; see O'Neill
1989, 99. Kant does remind us that parents try to have their children develop
a range of skills that will serve whatever ends they adopt down the road
(G 415).

[15] For a reconstruction of the argument that there must be "broad obliga-
tions," that is, imperfect duties, see O'Neill 1989, 230. The argument turns on
the claim that one does not have available principled ways of delimiting one's
obligation more tightly. Kant provides a subsidiary argument to support this
observation that is, however, less plausible than the observation itself is: any
further argument would establish a different ground of obligation. (DV 403;
compare A787f./B815f.)

[16] Loosely adapted from Herman 1993, chap. 6; compare Korsgaard 1996a,
98–100, Nell 1975, 79f.

[17] And more generally, as Cindy Stark has reminded me, not all violence is
an impediment to agency; *The Fight Club* is a recent film whose eponymous
institution can serve as an example.

[18] Whether this argument is to be understood as invoking a contradiction in
conception or a contradiction in the will depends on how contradiction in
conception is understood (see note 5, above), and on whether the former or
the latter of these problems is being emphasized. Herman presents it as invok-
ing your stake in your own agency generally, and so as exploiting a contradic-
tion in the will; on that reading, the duty not to avail oneself of violent means
would be imperfect. I want to leave to one side here the question of whether
we should be happy with an argument that makes abjuring violence out to be
an imperfect duty; what does matter is that the argument is endorsed by sev-
eral New Kantians, being treated by them—and this is a point for which I will
have a use later on—as something of an accomplishment.

[19] Herman (1993 53f.) gives an argument meant to block the following
objection: that you can avoid contradictions in the will by deciding to adopt
only projects guaranteed not to require resources provided by mutual aid, oth-
ers' talents, and so on.

[20] I'll provide arguments for these claims, as well as an explication of what is
meant by the "size" of an action, in section 6.

[21] There are other possible final clauses for the CI-maxim, for instance, "out
of respect for the moral law." That would express the point of the action, but
not perhaps an end. (I say "perhaps," because while such respect is not any-
thing like a goal, Kant redescribes respect for persons as treating them as
"ends in themselves," and persons are not, except in pathological cases, any-
thing like goals either; possibly the right gloss on Kant's term "end" would
make an end of respect for the moral law.) Another alternative final clause

might be "to promote the Highest Good"; I take it that the Highest Good is best understood as occupying the role, in Kant's account, of the formal end (but of course not the determining ground) of moral and rational action-in-general. For present purposes, however, it does not much matter what the final clause of the CI-maxim is taken to be, because final clauses do work at the contradiction-in-conception phase of the CI-procedure, and not at the contradiction-in-the-will stage. The argument I am about to construct will not need to demonstrate a contradiction in conception.

[22] Taking seriously the requirement that the CI-procedure is self-endorsing conforms to the approach taken by New Kantian metaethics. Korsgaard, for instance, extracts from a historical survey the lesson that self-endorsement is the only possible source of normativity; she subsequently generalizes the first-cut requirement, of a faculty's endorsing itself, to the general endorsement of all of one's faculties by all one's faculties (Korsgaard 1996b, 62, 65f.). On both the less and the more stringent versions of the view, a faculty's flunking itself out should strip the normativity from its pronouncements.

[23] *Need* the sort of maxim we have in mind be foundational? Perhaps you have some reason for a proposal as important as: always acting on the moral law (respect for persons, or a deep commitment to some set of values; see note 45). And normally, when you get around to running a maxim through the CI-procedure, you already think the course of action it proposes is a good idea, for some reason or other. But Kant himself seems to have thought that adherence to the Categorical Imperative could have no *further* ground; see note 39, below. That said, I don't think we have to settle the question here.

[24] Compare O'Neill 1989, 129f. But how could Kant have held both this view and accorded the importance that he did to imperfect duties? I think there really is a deep tension in his view here, but the parts of it that are on display are formally compatible. The imperfect duties are given as mandatory ends, not as maxims. Acting to promote a mandatory end requires formulating a more structured intention—a maxim—that specifies what, on a specified class of occasions, you will do to promote the end. (The class of occasions you specify may not exhaust the occasions on which you will, in one way or another, act to promote the end; those further occasions will have to be covered by other maxims.) Your maxim will then have to be checked for permissibility, using the universalizability test. For instance, to develop my talents is a mandatory end, but not yet a maxim. When I consider developing my talents by robbing banks, I now have a maxim, but this maxim fails at the contradiction-in-conception stage of the CI-procedure. (If everybody robbed banks, there would be no banks to rob.) So although I must find ways to develop my talents, I may not do it by robbing banks.

[25] Rawls tries to deal with a related objection—that moral restrictions will also get in the way of plans you may have—by appealing to a special class of interests, which he calls "true human needs"; the idea is that these are tied to different kinds of obstacles to agency (Rawls 2000, 173f.; such needs are an exegetical adaptation of his own notion of "primary goods"). So note that I am constructing the argument so that there is no plausible difference of this kind to which to appeal.

[26] Kant himself had been influenced by the Leibnizian tradition, and so there is some plausibility to a reading of the text on which permissible actions turn out to be unique and so required, and on which Kant's vision of the kingdom of ends is reminiscent of Leibnizian preestablished harmony. For now, we can say that the New Kantians find this sort of rigorism uncongenial, both because it is not in keeping with the *Zeitgeist*, and because the point of their project is to reclaim Kant as a predecessor to Rawlsian liberalism, a position that tries to maximize, rather than maximally constrict, freedom of action. We will provide a more principled reason for not falling back on rigorism below.

[27] Over and above that, you might be worried that your needing an exception shows that you did not have a (by Kantian lights) coherent intention in the first place. But since we cannot anticipate the exceptions our projects will require, insisting that a project is coherently intended only if it will not require exceptions would unacceptably undermine agency.

[28] G 423; we can imagine someone wealthy enough to pay for all the help he needs.

[29] Korsgaard 1999, 25, my emphasis. Quotations below are from pages 24–25 of this same paper.

[30] Compare Hill 1992, a paper whose title announces its subject to be a Kantian treatment of exceptions, and which turns out to be an attempt to work up more nuanced reasons for more nuanced—but formally universal and exceptionless—policies (in this case, "policies that involve taking the lives of terrorists," and possibly bystanders).

[31] One option that I have had suggested to me is that the burden be taken up by an Aristotelian sensitivity to when exceptions are called for. (That is, the more complex policy is to be implemented partly as the original, explicitly stated, but less complicated policy, and partly as a disposition to notice the specified classes of exception when they come up; Herman's "rules of moral salience" are a New Kantian attempt to integrate such patterns of attention and recall into the Kantian apparatus.) What matters for present purposes—and the reader should verify this as we walk through the argument—is that how the more complex rule is implemented makes no difference, first, to its social effects, and second, to the cognitive burden (computational and otherwise) involved in following it.

[32] Why can't the problem be solved by integrating the accumulating exceptions into cleanly formulated rules? After all, something like this goes on in the sciences; why not in ethics?

I think that the accretion of detail outpaces our ability to identify simplifying patterns; that's visibly true of such examples as the tax code. It's true in the sciences as well, and some philosophers of science, like Cartwright (1983), argue that the simplicity of the laws is consequently purchased at the expense of their truth. And there is a further obstacle in ethics. Recall that maxims have to capture the motivations the agent actually has. Those motivations are normally tied to a specific conceptual apparatus. Now, what we learn from the history of such simplifications in the sciences is that they are accompanied by fairly radical conceptual shifts. But we can insist on deploying a new set of concepts in one's maxims only if they can summon up corresponding motiva-

tional structure. As a matter of sociological fact, change in such concept-embedding motivations is very slow. So once again, we should expect the process of practical simplification to fall ever farther behind the accretion of legitimate exceptions to our rules. (For example, an anti-Semitic maxim will express a motivation tied to the concept "Jew." That means that we can't effect the gestalt or paradigm shifts that would render our maxims less complicated, when those shifts would involve replacing the concept "Jew," if we can't get rid of the anti-Semitism.)

[33] You may be inclined to look for more restricted versions of either the CI-maxim or the Revision Maxim; in that case, notice where the strategy lands you. The Kantian account needs machinery to prevent tailored versions of intentions to lie, cheat and steal from slipping through the CI-procedure. (See note 38, below, for a terse recapitulation of the problem.) So a restricted version of, say, the Revision Maxim needs to come with a *motivated* account of the machinery, and an argument demonstrating how, in light of that account, proposed revisions to the Revision Maxim are not blocked (even though unacceptable revisions to lying maxims and their ilk are). I don't have a tight argument showing that no such proposal can work—that's because it's hard to have a good enough sense of the range of possible proposals ahead of time—but I have no reason to expect that any will.

[34] Bendor 2003 reviews the history of (and obstacles to) the reception, in political science, of the methodological suggestion that human cognitive limitations need to be taken into account in theory construction.

Computer science has a well-established mathematical subdiscipline that studies computational complexity, and novice computer scientists are taught that intractability is ubiquitous. But it is remarkable that, half a century ago, as distinguished a founder of that discipline as Alan Turing (1950) could be quite certain that a program exhibiting intelligence would run on the hardware available in his day—that is, on what was, by our lights, scarcely a pocket calculator.

[35] See, for instance, Hayek 1989, at 143f., 149f., Hayek 1948, chaps. 2, 4, 7–9. Invisible-hand capitalists have taken to gloating over the vindication of Hayek's attack on socialism, but they may well have very similar problems; there is a growing body of evidence that market clearing is often a computationally intractable problem.

We can now give the principled response to the rigorist rejoinder broached in note 26. Even if there is a unique equilibrium point in the game of policy selection (and of course we have no proof that there is), we have no reason to believe that it is computationally accessible. If it is not, then it is not for practical purposes a solution at all. From the standpoint of practical reasoning, a solution that it is in principle impossible to produce is irrelevant.

[36] Some domains may look to be complicated, but be question-begging if used as counterexamples. I am arguing that humans have to negotiate their environments by deploying exceptions, rather than complicated rules. So one should not slide from thinking that, say, etiquette looks tricky, to thinking that, because tricky problems have to be handled by complicated rules, it displays a human ability to deploy complicated rules. In fact, etiquette is tricky

mostly because exceptions require a good deal of delicacy and judgment.

Recall that the CI-procedure is meant as a rational reconstruction. That means that we do not have to walk through the procedure aloud, before each action, but we must still possess the cognitive abilities needed to underwrite sensitivity to its dictates. Analogously, we do not write out parse trees before we utter sentences, but the grammatical complexity of the sentences we produce is limited by our cognitive capacities.

[37] Members of a maxim hierarchy are related to one another roughly as the answers to Anscombe's famous series of "Why?"-questions (1985, sect. 23); for work in the Anscombian tradition that takes up the ways in which "larger" actions embed and justify "smaller" actions, see Vogler 2003.

[38] In Herman, they are prominent as a partial solution to the so-called Problem of Relevant Descriptions. It is a familiar point that whether or not the CI-procedure rules any *actions* (as opposed to maxims) out of order depends on how much flexibility there is in selecting a maxim to test. Actions have indefinitely many descriptions, and one can always find a description of an action that really could be made the rule: lying may not be universalizable, but lying only to the naive and clueless probably is universalizable.

Herman has realized that the requirement that what is checked be the psychologically actual intention underlying the contemplated action, and not just any old description of it, will not produce a recognizably moral pattern of permissions (and, more important from a theoretical point of view, an *orderly* pattern) if agents' psychologies are sufficiently idiosyncratic. If the peculiarly tailored maxim (the one that slips past the CI-procedure) really *does* express your intent, then the CI-procedure will tell you that you can go ahead with it. The problem is not just that the CI-procedure usefully regulates only those with already very standard patterns of motivation, while giving the strange and the psychopathic *carte blanche*. It is that necessary but entirely unaccounted-for regimentation of agents' motivational structure has become the engine of the theory. This would amount to failure of the New Kantian theoretical enterprise, and Herman has devoted much thought and ingenuity to forestalling it (1993, esp. chap. 4; chap. 7, sect. 4; chap. 3, sect. 3.)

Maxim hierarchies can be used to provide a certain amount of anchoring against the tides of agential idiosyncrasy, because, for any given maxim that itself passes the CI-procedure, one can still test its superiors and inferiors in the hierarchy. (Something like this move can also be found in O'Neil 1989, 87, where "specific intentions" turn out to be "ancillary to more fundamental intentions *or principles* that might indeed have revealed moral unworthiness in the agent.") However, perhaps maxim hierarchies do not solve the Problem of Relevant Descriptions on their own: for all it has been argued, an agent could be thoroughly perverse, all the way up his casuistical hierarchy, in a manner that would slip through the CI-procedure.

[39] Kant does give a quick argument that looks like it should be working as a reductio, but that he claims instead establishes that "the first subjective ground of the adoption of moral maxims is inscrutable" (R 21, and esp. 21 n.). The problem is that the ground of the free adoption of a maxim must be sought, not in any incentive of nature, but in a further (freely adopted)

maxim; this generates a regress of maxims; but since humans have small-finite psychologies, they cannot support such a regress; their psychologically available reasons will bottom out, if you keep pressing, in a brute natural fact (typically an impulse or inclination).

In my view the best way to accommodate this tension is not to take refuge in "inscrutability," but to regard the demand that leads to embedding maxims in a hierarchy as just that: a demand, to make the effort to articulate one's motivations further when that is appropriate.

[40] The phrase is from O'Neill 1989, 87: "[i]f in welcoming my visitor with a cup of coffee I intentionally select a particular cup, my specific intention clearly cannot be universally acted on." Her distinction between "specific intentions" and "maxims" is intended to get you out of having to have such choices approved by the CI-procedure. See also her note 6 on that page, and 112, note 2.

[41] New Kantians *are* comfortable with judgment calls in the application of maxims; see, for instance, Nell 1975, 37. But they see that as a very different matter, having to do with the fringes of the theory, and not with its overall structure.

[42] Herman allows that "the maxims Kant uses are at all levels of generality" (1993, 220f.; she thinks her own account of maxims to be "unorthodox"). Korsgaard reminds us approvingly that "Kant proposes that we can tell whether our maxims should be laws by attending not to their matter but to their form" (1996b, 107), to which we can add that the size of a maxim is a matter of its matter and not of its form.

[43] O'Neill thinks of maxims as "underlying [practical] principles" (1989, 84f.), to be contrasted with both "aspects of action that are 'below the level of intention,'" and "our more specific intentions"; see 129, 151f., 158. In her earlier writing, the issue is "the amount of detail about an agent's circumstances and his proposed act which can be included in his maxims" (1975, 37).

[44] Not always: O'Neill suggests getting out of a class of hard cases not by refusing the title of maxim, but by simply waiving the test provided by the CI (1975, 76f.). Herman announces that "[e]ach interpreter of the CI must develop ways to set aside the puzzle maxims" (1993, 225).

[45] In the course of her efforts to address the Problem of Relevant Descriptions, Herman has endorsed reversing the characteristically Kantian priority of deliberative procedure to value, so as to let values underwrite and control the procedure (see Herman 1993, 153f., and compare C2 63); and you might think that some such appeal to values could be used to address the difficulties raised by the argument we have been developing. Herman has come to see the CI-procedure as expressing respect for persons (or "ends-in-themselves"), and this alternative way of motivating the CI-procedure allows one to forgo thinking of it as a consistency test. Making that move opens up the possibility of restricting the applicability of the CI-procedure, because an intention or policy (such as the policy of acting in accord with the CI-procedure) will no longer have to be considered practically inconsistent if it fails to pass the CI-procedure. (However, the alternative motivation does not fix what I claimed above was the decisive rejoinder to the Glass Ceiling Objections.)

This is radical surgery on the Kantian position that I hope can in the end be avoided, if only because it would make Kantian moral theory much less deep, and much less interesting, than it has the promise of being. To see how radical, recall the Kantian commitment to autonomy over heteronomy, and recall that whether one is autonomously or heteronomously related to one's evaluations is a matter of how they are accepted, and not a matter of their content. In principle, one's acceptance even of the importance of autonomy, or of the value of persons, could be heteronomous (for instance, if one believed that autonomy was supremely valuable because an authority had said so). Now, in Herman's revision of Kant's views, it is the value of persons and of autonomy that come first, and that underwrite allegience to the Categorical Imperative. So Herman is advocating a "Kantian" position on which one is to be heteronomous with respect to one's deepest and morally most central commitments.