THE BLACKWELL GUIDE TO

Kant's Ethics

EDITED BY THOMAS E. HILL, JR.

2009

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

5

Deriving the Supreme Moral Principle from Common Moral Ideas

Samuel J. Kerstein

In the Preface to the *Groundwork of the Metaphysics of Morals*, Kant sets out his goals: to locate and to establish the supreme principle of morality (G 4:392). He devotes Section I to the first goal. Working under the assumption that there is a supreme principle of morality, he tries to locate it in the sense of specifying its content.[1] Kant strives to find the supreme principle that, on reflection, we hold to be at work in our moral practice. His attempt rests on appeals to (what he takes to be) ordinary moral reasoning. Near the end of *Groundwork* I, he proclaims success. "Thus, then, we have arrived, within the moral cognition of common human reason, at its principle, which it admittedly does not think so abstractly in a universal form, but which it actually has always before its eyes and uses as the norm for its appraisals" (G 4:403–4). This principle is the Categorical Imperative. In *Groundwork* I, Kant's main concern is to show that if there is a supreme principle of morality then it is this imperative. It is not until *Groundwork* III that Kant tries to establish the Categorical Imperative, that is, to prove that we are all rationally compelled to conform to it.

This paper focuses on Kant's attempt to locate the Categorical Imperative – an attempt which, in the idiom of contemporary Kant scholarship, is called his "derivation" of this principle. In *Groundwork* I Kant derives one particular formulation of the Categorical Imperative, namely (a version of) his famous Formula of Universal Law: *act only on that maxim through which you can at the same time will that it become a universal law* (G 4:420–1).[2] The title of *Groundwork* I is "Transition from common rational to philosophic moral cognition." Kant's starting point is "common rational moral cognition," which is a fancy term for common moral ideas. How, from this starting point, does Kant derive the Formula of Universal Law? This paper sketches an answer to this question.

In my view, Kant's *Groundwork* I derivation of the Formula of Universal Law takes place in three main steps. First, Kant tries to pinpoint criteria that we, on

reflection, believe that the supreme principle of morality must fulfill. Second, Kant attempts to establish that no possible rival to the Formula of Universal Law fulfills all of these criteria. Third, at least implicitly Kant argues that the Formula of Universal Law remains as a viable candidate for a principle that fulfills all of them. With these three steps, Kant strives to prove that if there is a supreme principle of morality then it is this formula. In short, Kant argues by elimination. When we have before us a clear notion of the characteristics the supreme principle of morality must possess, Kant suggests, we are able to eliminate every candidate for this principle except the Formula of Universal Law (or equivalent principles). I call this interpretation of Kant's derivation the "criterial reading," since it emphasizes that Kant develops criteria that any viable candidate for the supreme principle of morality must fulfill.

The criterial reading of Kant's derivation is not the only one philosophers have proposed. (For alternatives, see, for example, Aune 1979, and Korsgaard 1996). Elsewhere, I have tried to demonstrate that the criterial reading compares well with other interpretations in terms of its textual plausibility and philosophical fruitfulness (Kerstein 2002). Here I limit myself to illustrating how the criterial reading helps us make sense of *Groundwork* I. In particular, I focus on how Kant appeals to "common rational moral cognition" in order to develop criteria for the supreme principle of morality.

The *Groundwork* I derivation culminates in the introduction of the Formula of Universal Law. In the sentence that immediately precedes its introduction, Kant poses a question: "But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it, in order for the will to be called good absolutely and without limitation?" (G 4:402). Kant is in effect asking which principle (law) can possess certain characteristics, namely ones that, according to ordinary moral thinking, the supreme principle of morality must possess. These characteristics are crystallized in the famous three "propositions" that Kant tries to establish in *Groundwork* I. The propositions encapsulate criteria for the supreme principle of morality. So by tracing how Kant arrives at his propositions, we will gain an understanding of how he develops these criteria.

The paper unfolds as follows. Parts II–III concern Kant's three propositions and the criteria for the supreme principle of morality implicit in them. The bulk of the discussion focuses on Kant's first proposition, namely that an action has moral worth just in case it is done from duty. Kant relies far more directly on ordinary moral thinking in defending this proposition than he does in defending the other two. Part IV illustrates how Kant might use criteria he develops in order to show that rivals to the Formula of Universal Law cannot be the supreme principle of morality. Unfortunately, these criteria do not enable him even to come close to eliminating all rivals. But Kant's argument gains in strength when we recognize that he suggests an additional criterion for the supreme principle of morality: it must be such that a plausible set of moral prescriptions (i.e., plausible relative to common rational moral cognition) would stem from the principle. Part V focuses on this additional criterion. According to the criterial reading, the final

step of the derivation is to show that the Formula of Universal Law remains as a viable candidate for a principle that fulfills all of the criteria, including that it generate a plausible set of moral prescriptions. Determining whether the Formula of Universal Law fulfills this criterion would obviously require delving into the details of how to interpret this principle. That is a project for another occasion. The final section (VI) of this paper is devoted to a concern one might have regarding the interpretation of the derivation offered here. The interpretation emphasizes the extent to which it depends on appeals to ordinary moral thinking. But how are such appeals, which seem to be appeals to experience, to be reconciled with Kant's view that the supreme principle of morality is an a priori principle?

Before focusing on *Groundwork* I and the criteria for the supreme principle of morality that Kant there suggests, it will be helpful to consider what he means by "supreme principle of morality" in the first place.

I

According to (what I call) Kant's basic concept, the supreme principle of morality would possess four characteristics. It would be practical, absolutely necessary, binding on all rational agents, and would serve as the supreme norm for the moral evaluation of action. I call this concept of the supreme principle of morality basic because Kant suggests it in the *Groundwork*'s Preface.

To say that a principle must be the supreme norm for the moral assessment of action suggests several things. The principle would obviously distinguish between morally permissible and morally impermissible actions as well as specify which actions are morally required. In addition, whether an action was morally good would depend on how it related to this principle. Kant implies, for example, that no action that violated the principle would count as morally valuable (G 4:390). Finally, as the supreme norm for the moral assessment of action, the supreme principle of morality would be such that all genuine duties would ultimately be derived from it (see G 4:421).[3] The supreme principle would justify these duties' status as such.

Kant says that the supreme principle of morality "must hold not only for human beings but for all *rational beings as such*" (G 4:408; see also G 4:389, CPrR 5:32). The supreme principle of morality would have an extremely wide scope: one that extended not only to all rational human beings, but to any other rational beings who might exist, for example, to God, angels, and intelligent extraterrestrials.

A third feature the supreme principle of morality would have to possess is that of being absolutely necessary (G 4:389). On every agent within its scope, for Kant every rational agent, the principle would hold without exception (G 4:408). For us human agents, the supreme principle of morality would be an unconditional command (i.e., a categorical imperative in one sense of the term). That we were obligated to perform the action it specified would not be conditional on our having any particular set of desires.

Finally, for Kant the supreme principle of morality must be practical, that is, a rule on account of which agents can act. Kant implies this in the *Groundwork* Preface by specifying that morally good actions involve an agent's acting for the sake of the moral law, that is, the supreme principle of morality (G 4:390). For Kant the supreme principle must be able to figure directly in an agent's practical deliberations.

Kant claims to find in ordinary moral thinking agreement that the supreme principle of morality would have to have at least some of these features. He suggests, for example, that, according to the "common idea of duty," a moral law "must carry with it absolute necessity" (G 4:389).

II

Now that we have an idea of what, according to Kant, the supreme principle of morality would have to be like, let us focus on the details of his attempt to show that if there is such a principle then it is the Formula of Universal Law. As I indicated, Kant's attempt involves his developing three propositions, each of which implies a criterion for the supreme principle of morality. It makes sense to begin with Kant's first proposition. Although Kant does not explicitly state it, it is widely, and, I believe, correctly taken to be the following: An action has moral worth if and only if it is done from duty.

As a first step towards understanding this proposition we need to delve briefly into Kant's famous discussion of a good will. For Kant tells us that the concept of duty "contains that of a good will though under certain subjective limitations and hindrances" (G 4:397). Let us focus on Kant's discussion of the good will as it relates to us, agents who can indulge their inclinations and thereby act contrary to what morality requires. In this context, Kant seems to use the notion of a good will in two ways. According to the first usage, a good will is a particular sort of *willing* or, what for him amounts to the same thing, of acting. Kant writes of "the unqualified [*uneingeschränkten*] worth of actions" (G 4:411), presumably of actions done from duty. For he has earlier stated that actions from duty have "unconditional and moral worth" (G 4:400). Since, according to Kant, the good will is good without qualification [*ohne Einschränkung*], it appears that sometimes "good will" refers to a certain kind of action, that is, action done from duty.

According to a second usage, "good will" refers not to a particular kind of action an agent might perform but rather to a kind of character she might have. An agent has a good will in this usage, I believe, just in case she is committed to doing what duty requires, not just in this or that particular action, but overall. If an agent has this commitment, then she will presumably sometimes act from duty. (For example, she will invoke duty as her incentive to do what is morally required in cases in which she is tempted by her inclinations to act contrary to what morality demands.) In the first paragraph of *Groundwork* I, Kant intimates that having a good will amounts to having a certain kind of character. Right after suggesting

that a good will is good without qualification, he tells us that certain qualities of temperament, for example, courage or resolution, "are undoubtedly good and desirable for many purposes, but they can also be extremely evil and harmful if the will that is to make use of these gifts of nature, and whose distinctive constitution is therefore called *character*, is not good" (G 4:393).[4] Sometimes Kant employs what we might (following Karl Ameriks 1989, pp. 54–9), call the "whole character" conception of a good will. This is not the place to consider in detail how these conceptions relate to one another. But I suspect that in Kant's considered view the only way an agent can have a good will in the first (particular action) sense is if he has a good will in the second (whole character) sense. In other words, Kant holds that only an agent who has an overall commitment to doing his duty can act from duty on any particular occasion.

In any case, the particular action notion of a good will is more important for our purposes. Kant suggests that a good will in the sense of good *willing* is equivalent to acting from duty. And, according to "common understanding" (G 4:394), this willing (or, equivalently, acting) has a special, moral worth. First, it is unconditionally good. In all possible circumstances in which it appears, a good will is good; moreover, the degree of its goodness does not vary according to its effects. Even if a good will "should wholly lack the capacity to carry out its purpose – if with its greatest efforts it should yet achieve nothing and only the good will were left . . . then, like a jewel, it would still shine by itself, as something that has its full worth in itself" (G 4:394). Second, according to ordinary moral thinking, the worth of a good will is especially high, Kant claims. We take a good will (good willing) to be preeminently valuable (see G 4:394 and G 4:401). That presumably implies that, in our view, no particular action that is not done from duty is as valuable as any action that is done from duty.

Let us now return to Kant's first proposition. It states that an action has moral worth if and only if it is done from duty or, equivalently, that all and only actions done from duty have moral worth. The two key concepts in this proposition are obviously those of moral worth and of acting from duty. Moral worth, as we just noted, is unconditional and preeminent worth. At this stage in his argument, Kant does not explain precisely what acting from duty amounts to. But from the *Groundwork*'s Preface, it's easy to discern the basic idea he has in mind (and takes his reader to have in mind as well). Acting from duty is doing something "for the sake of" the moral law (G 4:390). In other words, to act from duty is to do something because a valid moral principle (or at least a principle one takes to be valid) prescribes that one do it. A more rigorous account of acting from duty emerges from Kant's discussion of his third proposition.

In the *Groundwork*, Kant sets out grounds for rejecting the notion that actions from motives other than duty have moral worth. Yet he apparently finds it unnecessary to argue that all actions done from duty possess such worth. Consider, for example, his discussion of self-preservation. Kant suggests that we have a duty to preserve our life and that, the vast majority of the time, when we take steps to preserve it we are acting from an immediate inclination to stay alive. "But on this account," Kant says, "the often anxious care that most people take of [their life]

still has no inner worth and their maxim has no moral content. They look after their lives *in conformity with duty* but not *from duty*" (G 4:397–8). Kant takes it to be obvious that if a person preserves his life not from inclination but from duty, "his maxim has moral content," and thus acting on it has moral worth. He assumes that "common rational moral cognition" needs no coaxing in order to see that actions done from duty possess moral worth.

In contrast, Kant does think we need a bit of help in order to discern that *only* actions from duty have moral worth. He highlights two conditions on actions with such worth, both of which he takes to be accepted by common rational moral cognition. He then intimates that no action from inclination could meet these conditions. Kant introduces the first condition in the *Groundwork* Preface:

> in the case of what is to be morally good, it is not enough that it *conform* with the moral law; but it must also be done *for the sake of the law*; without this, that conformity is only very contingent and precarious, since a ground that is not moral will indeed now and then produce actions in conformity with the law, but it will also often produce actions contrary to the law. (G 4:390)

Morally valuable action, Kant here suggests, is action done from a motive that will not produce actions contrary to duty. In the *Groundwork*, Kant maintains that acting "for the sake of the law," that is, doing something because you take it to be required by moral principle, meets this condition, while acting from inclination does not.

Kant invokes this condition in his famous discussion of the "philanthropist" (or "friend of humanity") (G 4:398). Before the discussion begins, Kant suggests a distinction between acting from a mediate inclination (self-interest) and acting from an immediate inclination (G 4:397). A mediate inclination to do something is an inclination to do it for the sake of fulfilling some further inclination. The shopkeeper in Kant's example presumably has a mediate inclination to charge his customers fairly. He wants to do it, but merely as a means to satisfying another end, for example, that of having a thriving business. An immediate inclination to do something is an inclination to do the thing itself. Since he is "sympathetically attuned," the philanthropist presumably has an immediate inclination to promote the well-being of others. His inclination to help them is not one that he strives to satisfy merely to fulfill some further desire. Kant, of course, denies that acting from this inclination has moral worth. Doing so, he says, is like acting from other inclinations, for example, the inclination to honor, "which, if it fortunately lights upon what is in fact in the common interest and in conformity with duty and hence honorable, deserves praise and encouragement but not esteem" (G 4:398; see also R 6:30–1). Here Kant underscores the possibility that in acting from an immediate inclination to help others, that is, from sympathy, an agent might do something that conflicts with duty. To echo a well-known example (Herman 1993, pp. 4–5), someone might, because of his sympathetic temperament, have an immediate inclination to help someone he sees late one night hurriedly struggling to move

a sculpture out the back door of an art museum and into his waiting car. Since the philanthropist is acting from an immediate inclination, and thereby doing something that might fail to accord with duty, his action, Kant suggests, does not have moral worth.

Yet in his discussion of the philanthropist Kant points to a further condition he places on an action's having moral worth (Herman 1993, pp. 5–6). Kant says that the maxim on which the philanthropist acts "lacks moral content, namely that of doing such actions not from inclination but *from* duty" (G 4:398). Kant does not tell us explicitly what the philanthropist's maxim is. From the description Kant provides, however, we can assume that it is something like the following: "Because I want to help others, I will promote their happiness." This maxim, says Kant, lacks moral content, and it is not hard to pinpoint a reason why. The maxim reflects no commitment to the action's being morally permissible, that is, in accordance with what moral principle requires. In other words, the maxim expresses no interest in the rightness of the kind of action it specifies, namely promoting others' happiness. If we reflect on our ordinary moral understanding, suggests Kant, we find that we are willing to attribute moral worth only to actions done on maxims that (if fully specified) reflect a commitment to doing only what is morally permissible. The grounds of a morally valuable action, that is, its motive, must express an interest in the action's moral rightness. This is Kant's second condition for an action's having moral worth.

It is a necessary condition, not a sufficient one, that when an agent does some particular thing, he is committed to its being morally permissible, that it does not entail that his action has moral worth. What the agent does might be morally permissible, but not morally required. And for Kant only morally required actions can have moral worth. According to Kant, of course, actions from duty fulfill this second condition. In them, an agent's basis for acting, that is, his maxim, obviously expresses concern for his action's moral rightness, for it invokes the notion that actions of its kind are morally required.

Kant would insist that an action might fulfill the first condition for moral value without fulfilling the second (Herman 1993, p. 5). Suppose, for example, that the philanthropist's immediate inclination to help others were such that it served as the basis only for morally permissible actions. In that case, the philanthropist's beneficent actions would fulfill Kant's first condition: they would be done on a motive that always produced actions conforming to duty. Nevertheless, the philanthropist's actions would still not have moral worth, for the grounds of his actions would fail to express concern for their moral rightness, thereby running afoul of the second condition.

Kant's first proposition and his defense of it have attracted ample critical attention. Kant is perhaps too quick to conclude that, according to common rational moral cognition, an action has moral worth only if it fulfills his two conditions (and thus only if it is done from duty). He might also be precipitous in assuming widespread endorsement of the notion that all actions from duty have moral worth. My own view (Kerstein 2002, pp. 114–38) is that Kant is on much stronger ground in claiming that, according to common rational moral cognition, all

actions from duty have moral worth than he is in claiming that only actions from duty have moral worth. In any case, Kant's main appeals to ordinary moral thinking occur in his discussion of the special value possessed by a good will, as well as in the closely related discussion leading up to his first proposition.

III

The arguments Kant suggests for the second and third propositions are far less directly tied to intuitive moral judgments than his arguments for the first. In his "second proposition," Kant says that "an action from duty has its moral worth *not in the purpose* to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of the action but merely upon the *principle of volition* in accordance with which the action is done" (G 4:399–400). Later Kant says that "the moral worth of an action does not lie in the effect expected from it" (G 4:401; see also G 4:435).

Kant here invokes the notion of a principle of volition or maxim. We've already made use of this notion, but it makes sense to pause here to get a more precise idea of what a maxim is. The brief account of maxims that follows is certainly not the only plausible one, but it will serve to fix ideas. A maxim is a "subjective principle of acting" (G 4:421n; see also G 4:400n). It is a subjective principle in that it is held by some agent, it can be freely adopted or discarded by her, and it applies only to her own actions. An agent's maxims are principles of acting in that they play a role in the generation of her actions. An agent acts on maxims. When fully specified, a maxim includes a description of a kind of action to be performed in a kind of situation, as well as a specification of the agent's end and his incentive in performing it. An example of a maxim is the following: "From self-love, during my free time I exercise in order to stay in shape." (Self-love is the agent's incentive; staying in shape is her end.) According to Kant, whenever an agent acts, she does so on some maxim, even though she might not have it explicitly in view.

Kant's second proposition says essentially that an action done from duty derives its moral worth from its maxim rather than from its effects. The proposition relies on a distinction between an action (which is always done on some maxim) and its effects. For Kant, to act is to exercise one's will (Kerstein 2002, pp. 20–1). It is to try, based on some principle of volition, to realize a state of affairs (an object or end). This state of affairs (or whatever state of affairs actually results from the action) is an effect of the willing. Acting consists in the willing itself, not in its effects (see G 4:400). According to the second proposition, it is the maxim behind an action done from duty that gives it moral value, rather than the action's results.

Implicit in *Groundwork* I is a straightforward argument for the second proposition. Suppose that, contrary to it, the moral worth of an action from duty *did* stem from its effects. There would, then, be possible circumstances in which an action from duty did not have moral worth, namely, ones in which the action

failed to produce certain effects. For Kant, however, if an action is done from duty, then it has moral worth, no matter what the circumstances may be. His first criterion incorporates this view. Moral worth is "unconditional," Kant suggests (G 4:400). Therefore, as the second proposition indicates, the moral worth of an action from duty does *not* stem from its effects. For example, suppose that an agent holds the supreme principle of morality to be: "Always do what you believe will please God." Moreover, contrary to the second proposition, the agent maintains that the moral worth of her conforming to this principle because the principle requires it, that is, the moral worth of her acting from duty stems from its effects. Whether her action has moral worth, she thinks, depends on whether it actually pleases God. Since, as a fallible being, she might be mistaken as to what would please God, there would presumably be possible circumstances in which her acting from duty would not actually please her/him. In these circumstances, the agent would be compelled to maintain, her acting from duty would be devoid of moral worth. But this acknowledgment would contradict Kant's first proposition, according to which a sufficient condition for an action's having moral worth is that it be done from duty. In short, Kant defends the second proposition by appealing to the first. That the effects of our actions can give them "no unconditional and moral worth," he says, "is clear from what has gone before" (G 4:400). What has gone before, of course, is Kant's discussion of the relations between acting from duty and moral worth: a discussion, based on common rational moral cognition, that lays the basis for his first proposition.

According to Kant's third proposition, "duty is the necessity of an action from respect for law" (G 4:400, emphasis omitted). This proposition fills in some details regarding what it means for an action to be done from duty. According to the proposition, if an action is done from duty then what determines it is "objectively the *law* and subjectively pure respect for this practical law, and so the maxim of complying with such a law even if it infringes upon all my inclinations" (G 4:400–1). By "law" here, Kant means a universally binding and absolutely necessary practical principle. When an agent acts from duty, Kant here suggests, his action stems from the notion, which is incorporated into his maxim, that a practical law requires it. Kant even says that "an action from duty is to put aside entirely the influence of inclination" (G 4:400). So, in his view, an agent who needs to rely on an inclination in order to get something done fails to act from duty. If an agent acts from duty, the notion that a law requires her action itself generates enough motivation for her to do it. It generates this motivation, Kant suggests, at least in part by producing in her a feeling of respect for the law. Kant develops his concept of respect in detail in the *Critique of Practical Reason* (CPrR 5:71–89). It is very complex, and we have no need to explore it here. But we do need to hold in view that, according to Kant's third proposition, when an agent acts from duty, her notion that her action is required by a practical law provides her with sufficient motive for doing it. In other words, this notion gives her a ground sufficient to determine her will.

But how does Kant defend this proposition? He suggests, but does not explicitly make, the following argument.[5] Suppose that in an action done from duty the

notion that the action was required by a practical law did *not* give an agent sufficient motive to perform it. In that case, Kant suggests, the additional motive necessary for the agent to perform the action would have to be the agent's expectation that her action would bring about certain effects (G 4:401). But now further suppose that the action did not produce the expected effects. In that case, the agent would be rationally compelled to agree that the action had *less value* than it would have had if the expected effects had come to fruition. After all, if, in the agent's view, the action's value was not at all contingent on the effects being produced then why would she need to acquire part of her motivation for doing it from the prospect that the effects would be produced? But if an action done from duty has less value than it otherwise would have as a result of its not producing certain effects, then its value is not unconditional. And this result conflicts with Kant's first proposition, according to which all actions from duty have moral, and thus unconditional, worth. The result also conflicts with his second proposition, since according to it the moral worth of an action does not depend (at all) on the action's effects. So it makes sense for Kant to suggest, as he does (G 4:400), that his third proposition follows from the previous two.

Kant's main aim in articulating his three propositions is to derive the supreme principle of morality, that is, to show that if there is such a principle, then it is the Formula of Universal Law. Each one of the propositions implies a corresponding criterion that the supreme principle of morality must fulfill. The first proposition says that an action has moral worth if and only if it is done from duty. According to the criterion implicit in this proposition, the supreme principle of morality must be such that all and only actions conforming to it because the principle requires it, that is, all and only actions done from duty, have moral worth. The second proposition says that an action done from duty derives its moral worth from its maxim rather than from its effects. So whatever the supreme principle of morality is, goes the second criterion, the moral worth of conforming, from duty, to it must stem from the maxim of the action, not from its effects. According to the core of the third proposition, when an agent acts from duty, her notion that her action is required by a practical law provides her with sufficient grounds for acting. The criterion implicit in this proposition is the following: the supreme principle of morality must be such that an agent's notion that it is a practical law and that it requires her to do something gives her sufficient motive to do it. It's up to the agent, of course, whether she acts on this motive and does what is required rather than, say, indulging an inclination to do something else.

Let us again note that in the sentence preceding his initial presentation of the categorical imperative, Kant asks: "But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it, in order for the will to be called good absolutely and without limitation?" (G 4:402). He is, in effect, asking what law (principle) can fulfill each of these three criteria for the supreme principle of morality: the third criterion, which invokes an agent's representation of a law as a sufficient motive for her action; then the second criterion, which incorporates the notion that the moral worth of an action does not stem from its effects; and finally the first criterion,

which specifies when an action, that is, an instance of willing, has moral and thus unconditional worth. If we can show that a particular principle is unable to fulfill any one of these criteria, then we can, Kant suggests, eliminate it as a viable candidate for the supreme principle of morality. If Kant's derivation of the categorical imperative is successful, then we should be able to see that the only principle that remains as a viable candidate for satisfying all three of these criteria (plus those criteria implicit in Kant's basic concept of the supreme principle of morality) is the Formula of Universal Law (or an equivalent principle).

IV

Unfortunately, it would be unduly optimistic to say that we are in position to see this. Kant moves extremely quickly from the criteria he develops to the conclusion that the only viable candidate for fulfilling all of them is the Formula of Universal Law. He seems to leave it to us to fill in the details regarding precisely how rivals get eliminated. Even if, employing Kant's criteria, we eliminate all rivals that come to mind, it is not clear how we can be confident that we have not overlooked some other rival.[6] Nevertheless, we are well-situated to see how we might use Kant's criteria to dismiss some well-known principles as viable candidates for the supreme principle of morality.

Kant does not explicitly argue against utilitarianism. But let us consider a utilitarian principle, U: "Always perform a right action: one that yields just as great a sum total of well-being as would any alternative action available to you." Let us suppose, as it seems reasonable to do, that the utilitarian embraces this principle largely on the grounds of her being convinced of the following. First, the amount of goodness in the world depends solely on the sum total of individual well-being in it – the higher the sum total, the more goodness. Second, the rightness of an action depends solely on the goodness of its consequences. More precisely, an action is right if that which results from it is at least as good as that which would have resulted from each of the alternative actions available to the agent.

Although U derives from these un-Kantian convictions, it would be precipitous to dismiss it as a candidate for the supreme principle of morality on the grounds of a manifest failure to conform to Kant's basic concept of this principle. U could, it seems, be a practical, absolutely necessary, universally binding, fundamental norm for moral evaluation of action.

But U runs afoul of Kant's further criteria for the supreme principle of morality. The utilitarian might insist that an agent can, from duty, comply with U. After all, what would prevent her from performing a right action just because U commands her to do so? Yet she is committed to the following view: whether an agent's conforming to U from duty has moral worth depends solely on the action's effects, specifically its effects on well-being. For she holds that the amount of goodness in the world (including the "moral worth" of actions) depends *solely* on the amount of well-being in it. So the utilitarian cannot, rationally speaking, maintain that U fulfills Kant's second criterion, namely that the supreme principle of

morality be such that the moral worth of conforming to it from duty stems not at all from that action's effects.

If we think of a consequentialist principle as one according to which the goodness of each and every action depends to some extent on the action's effects (in addition to the "effect" that the action has taken place), then it is easy to show that no consequentialist principle fulfills Kant's second criterion. For even the staunchest proponent of such a principle would have to acknowledge that she is committed to the view that the value of acting from duty depends at least in part on what that action produces.

So based ultimately on an appeal to the notion that, according to ordinary moral thinking, actions from duty have a special worth, Kant develops three criteria for the supreme principle of morality. Assuming these criteria are sound, Kant has solid grounds for dismissing some of the Categorical Imperative's rivals for status as viable candidates for the supreme principle of morality.

V

However, if, as a basis for dismissing rivals, Kant has only these three criteria, coupled with those implicit in Kant's basic concept of the supreme principle of morality, he is vulnerable to a serious criticism. Using these criteria, he would be helpless to eliminate rivals that, one would think, would have almost no chance of being the supreme principle of morality. Consider the bizarre principle, BP: "Act only on that maxim such that you *cannot*, at the same time, will that it become a universal law." Assuming that the Categorical Imperative could be a universally valid, absolutely necessary, supreme practical principle, why couldn't BP be such a principle? What argument does Kant have at his disposal that would show it to be impossible for BP to have these characteristics? Moreover, it seems that a proponent of BP would be able consistently to maintain that an action has moral worth if and only if it is done because BP requires it, that such an action's moral worth would not stem from its effects, and so forth. He would not be rationally compelled to acknowledge that BP runs afoul of the criteria implicit in Kant's three propositions.

Another, less provocative, example of a principle Kant would be unable to dismiss on the basis of his criteria is the following principle of weak universalization, WU: "Act only on that maxim which, when generalized, could be a universal law." WU is not equivalent to the Formula of Universal Law. And Kant himself suggests that a maxim of non-beneficence could, when generalized, constitute a universal law (G 4:423). Since a world where no one acted beneficently is indeed a coherent possibility, acting on a maxim of non-beneficence does not violate WU. On Kant's view, of course, acting on such a maxim does run afoul of the Formula of Universal Law. It does so, he thinks, because as a rational agent it is not possible to act on it and, *at the same time*, will that its generalization be a universal law. On the basis of the criteria discussed thus far, Kant does not appear to have the tools to eliminate WU as a contender for

the supreme principle of morality. For not only is it possible that WU satisfies Kant's basic concept of the supreme principle of morality, but there seems to be no reason to think that it couldn't fulfill the criteria suggested by his three propositions.

In my view, this difficulty prompts us to see that Kant actually suggests one further criterion for the supreme principle of morality. It must be such that a plausible set of duties, that is, plausible relative to common rational moral cognition, would stem from the principle. Both BP and WU could be eliminated through an appeal to this criterion. According to ordinary moral thinking, contrary to BP and to WU, we have a duty of beneficence.

A textual basis for this criterion is not hard to discern. In *Groundwork* II, Kant offers a derivation of the Formula of Universal law that parallels his derivation in *Groundwork* I. Right after he arrives at this formula, Kant says: "Now, if all imperatives of duty can be derived from this single imperative as from their principle, then, even though we leave it undecided whether what is called duty is not as such an empty concept, we shall at least be able to show what we think by it and what the concept wants to say" (G 4:420–1). The derivation is not complete unless "all imperatives of duty" can be derived from the imperative Kant proposes as the only viable candidate for the supreme principle of morality. By "all imperatives of duty," Kant apparently means all imperatives that we, reflective rational agents, take to express our moral duties. Kant proceeds, of course, to try to show that four such imperatives, including, for example, a requirement not to make false promises for financial gain, follow from the Formula of Universal Law. He then says: "These are a few of the many actual duties, *or at least of what we take to be such*, whose derivation from the one principle cited above is clear" (G 4:424, emphasis added). If these duties' derivation from the Formula of Universal Law were not clear, for example, if it simply did not follow from the formula that we had them, then, Kant implies, we could not accept this formula as the only viable candidate for the supreme principle of morality. In the short paragraph (G 4:420–1) following his statement of the Formula of Universal Law, Kant indicates an important criterion for any viable candidate for the supreme principle of morality. We must be able to see how it follows from this candidate that, if it were established, we would indeed have moral duties that we are convinced we do have. (For further textual evidence that Kant embraces this criterion, see Kerstein 2002, pp. 87–9.)

With this additional criterion in place, Kant can advance towards eliminating rivals for status as viable candidates for the supreme principle of morality. But in order for his derivation of the Formula of Universal Law to succeed, he would need not only to show that rivals are unfit to satisfy the criteria he indicates, but also that it remains viable to think that his candidate can satisfy them. Yet this latter task poses serious challenges. Kant offers various, supposedly equivalent, formulations of the Categorical Imperative. For example, in addition to the Formula of Universal Law, he offers the Formula of Humanity: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (G 4:429, emphasis omitted).

Both of these principles are difficult to interpret; it's far from obvious what either one would require us to do. In my view, it is very unlikely that the Formula of Universal Law would generate a set of duties acceptable to ordinary moral thinking (Kerstein 2002, pp. 168–174). The Formula of Humanity seems more promising on this score, but on interpretations suggested recently by Kantians, it too has implications that fail to square with common notions of morality (Kerstein 2002, pp. 177–187).

In sum, in *Groundwork* I Kant offers a derivation of the Formula of Universal Law. He tries to show that if there is a supreme principle of morality then it is this principle (or its equivalent.) He suggests a three-step process towards attaining this goal, I believe. The first step is to develop criteria for the supreme principle of morality. In order to do this, Kant appeals at key points to common rational moral cognition. He bases his notion that supreme principle must have a scope extending to all rational beings on such an appeal. The criterion implicit in his first proposition, as well as the criteria implied in his second and third propositions, stem ultimately from his notion that, according to ordinary moral thinking, an action has moral worth if and only if it is done from duty. And the last criterion we discussed appeals directly to the moral verdicts of common sense. The Formula of Universal Law is, of course, not the only principle that philosophers have sought to elevate to the status of the supreme principle of morality; Kant's principle has many rivals. The second step in the derivation is to eliminate these rivals on the basis of their manifest inability to fulfill all of the criteria. The third step is to show that the Formula of Universal Law remains as a viable candidate for fulfilling all of the criteria. Showing this involves demonstrating that this formula indeed generates a plausible set of duties relative to ordinary moral thinking.

Kant does not have to prove that the Formula of Universal Law actually does fulfill the whole set. For that would entail establishing that this principle is absolutely necessary and universally valid. And that is a project that Kant undertakes in the notoriously difficult third chapter of the *Groundwork*. Each one of these steps is controversial. But, in my view, Kant nevertheless offers a coherent, philosophically interesting argument for his conclusion that either there's no supreme principle of morality or it is the Formula of Universal Law.

VI

Let me close by considering an objection to this account of Kant's derivation. The objection is that it is not consistent with Kant's claim that the supreme principle of morality must be an a priori principle. In particular, consider the criterion according to which the supreme principle must be capable of generating duties that cohere with the moral duties we take ourselves to have. Does not whether we conclude that a given principle meets this criterion rest on experience, that is, our particular experience of morality? Already in the *Groundwork* Preface Kant says that the ground of an obligation to conform to the supreme principle of morality must be sought "a priori simply in concepts of pure reason" and that any principle

that "rests in the least part on empirical grounds, perhaps only in terms of a motive, can indeed be called a practical rule but never a moral law" (G 4:389).

In order to respond to this objection, we need to understand two senses in which, according to Kant, the supreme principle must be an a priori rather than an empirical principle. It must be a priori in both (what I call) a motivational sense and an epistemological sense.[7]

Beginning with the former, the supreme principle of morality must be such that all rational agents always have available to them a sufficient motive for abiding by it. (Whether they actually act on this motive or some other one, such as an inclination, is another question.) But that means that their having sufficient motive available to them to conform to the principle must not depend on anything empirical, that is, on their particular inclinations or even on their nature, insofar as this nature is not necessarily shared with all rational agents. A principle is a priori in the motivational sense just in case any rational agent's having available to him a sufficient motive for abiding by it is not conditional on anything empirical. (In effect, a principle is a priori when it fulfills the criterion implicit in Kant's third proposition). A practical principle would be empirical, for example, when a rational agent's having sufficient motive to abide by it was conditional on his expectation that abiding by it would give him pleasure (CPrR 5:9n).

Kant's appealing to experience in his derivation of the Formula of Universal Law does not seem incompatible with all rational agents having an empirically unconditioned motive at their disposal for abiding by this formula. That we rely on our moral experience in pinpointing the supreme principle of morality does not, for example, seem to entail that our having at our disposal sufficient motive to comply with it is conditional on our expectation that doing so will get us something we want.

The second sense in which, according to Kant, the supreme principle of morality must be a priori is what I call the epistemological sense. Kant states that a practical law, and thus the supreme principle of morality, must be knowable a priori (CPrR 5:26; see also G 4:440). In the *Critique of Pure Reason*, Kant defines a priori knowledge as "knowledge absolutely independent of all experience" (CPR 3:3). If we had a priori knowledge of a judgment or proposition, this knowledge would have to be "absolutely independent" of all experience in the following sense: it would have to be *grounded* or *legitimated* without appeal to any particular set of experiences (see Allison 1983, p. 78). So, it seems, if we had a priori knowledge of the supreme principle of morality, that is, if we knew that it was necessarily binding on all rational agents, this knowledge would likewise have to be grounded or legitimated without appeal to any particular set of experiences.[8] This interpretation gains support from the *Groundwork*. After telling us that a moral law must be binding on all rational agents, Kant claims that "the ground of obligation here must not be sought in the nature of the human being or in the circumstances of the world in which he is placed, but a priori simply in concepts of pure reason" (G 4:389).

Does the account sketched above of Kant's *Groundwork* I derivation clash with the notion that the supreme principle of morality must be a priori in this sense? I

do not believe so. But before explaining why, let me begin with a blunt claim. It would be a mistake to maintain that in *Groundwork* I Kant proves (or could even reasonably take himself to prove) a priori that if there is a supreme principle of morality, then it is the Formula of Universal Law. Maintaining this would be a mistake even for those who reject the notion that, according to Kant, a criterion for the supreme principle is that it generate duties in line with those that we take ourselves to have. For Kant's first proposition, which is a cornerstone of the derivation on any plausible interpretation, is based largely on an appeal to ordinary moral thinking. Kant does not establish a priori that an action has moral worth when it is done from duty. To the extent that he argues for this proposition, he does so at least in part by appealing to our reactions to a range of cases, such as those involving the philanthropist. Kant himself announces his starting point in *Groundwork* I to be "common cognition," which amounts to ordinary, everyday reflection on things moral (G 4:392). Kant appeals to particular sets of experience, not merely to concepts of pure reason, in his derivation of the Formula of Universal Law.

That is not surprising; for Kant does not assert that his *Groundwork* I derivation rests solely on a priori grounds. What he does claim is that the supreme principle of morality must be knowable a priori. It must be possible to have a certain kind of knowledge that it is valid, namely knowledge that is not based on appeals to any particular set of experiences. The derivation's being based partly on appeals to our moral experience does not itself undermine this claim. In *Groundwork* I we reflect on our judgments regarding particular sorts of cases in order to see that a certain principle is at work in our moral practice. This does not entail that we cannot know this principle a priori.

In *Groundwork* III, Kant attempts to establish the Formula of Universal Law (or at least a principle closely resembling it). He tries to show that it would be irrational for any being within its scope, that is, any rational agent, to fail to comply with it. Any argument that proved this, Kant believes, could not be based on appeals to experience. The argument of *Groundwork* III is difficult to pinpoint. But if we assume that the argument does not (even indirectly) appeal to experience, then the following point becomes evident. In establishing the Formula of Universal Law, Kant would, in effect, show that it is knowable a priori. For he would show something stronger, namely that it is known a priori, at least by those who understand the argument.

Notes

1 Kant's main task in *Groundwork* II also seems to be to derive the supreme principle of morality – in all the complexity of its various formulas.

2 Strictly speaking, the principle is a preliminary version of the Formula of Universal Law, namely: "*I ought never to act except in such a way that I could also will that my maxim should become a universal law*" (G 4:402).

3 In *Groundwork* II, Kant says that the "the categorical imperative," the principle he takes to be the supreme principle of morality, is "the canon of moral appraisal of action in general" (G 4:424). On the next page (G 4:425), Kant says: "we have . . . set forth distinctively and as determined for every use the content of the categorical imperative, which must contain the principle of all duty (if there is such a thing at all)."

4 Later Kant is discussing a man who is by temperament cold and indifferent to others, but who, from duty, acts beneficently. "It is just then," says Kant, "that the worth of character comes out, which is moral and incomparably the highest" (G 4:398–9). This passage suggests that "good will" refers not merely to a particular kind of action, but to a kind of character that can be expressed in action.

5 In my view, Kant suggests this argument at (G 4:401). However, it would also be reasonable to consider the argument to be a reconstruction rather than an interpretation of this stretch of the derivation.

6 Kant does offer a table that is supposed to give an exhaustive classification of rival moral principles. But it is questionable whether this table is complete. See Kerstein 2002, pp. 140–4.

7 For a different account of Kant's emphasis on the a priori in the development of his moral philosophy, see Hill 2002.

8 Why does Kant say that the supreme principle of morality must be knowable a priori? The supreme principle of morality would have to be unconditionally and universally valid, thus admitting of no possible exception. But in Kant's view if a principle can be justified only by appeal to particular experiences then it cannot be known that no exception to it is possible. That experience has thus far shown that there is no exception to a principle fails to entail that there will be none. To bring the point to the issue at hand, that experience has thus far shown that a given principle generates all the duties we take ourselves to have does not entail that the principle will always generate all these duties. For it to be known that there can be no exception to a principle, the principle's validity must be grounded a priori.

Bibliography

Allison, H. 1983: *Kant's Transcendental Idealism.* New Haven: Yale University Press.

Ameriks, K. 1989: Kant on the good will. In O. Höffe (ed.), *Grundlegung zur Metaphysik der Sitten; Ein kooperativer Kommentar.* Frankfurt am Main: Klostermann.

Aune, B. 1979: *Kant's Theory of Morals.* Princeton: Princeton University Press.

Herman, B. 1993: *The Practice of Moral Judgment.* Cambridge, MA: Harvard University Press.

Hill, T. E., Jr. 2002: *Human Welfare and Moral Worth.* Oxford: Oxford University Press.

Kerstein, S. 2002: *Kant's Search for the Supreme Principle of Morality.* Cambridge: Cambridge University Press.

Korsgaard, C. 1996: *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press.