Groundwork for the Metaphysics of Morals

Edited by

Christoph Horn and Dieter Schönecker

in cooperation with

Corinna Mieth

2006

Walter de Gruyter · Berlin · New York

Samuel J. Kerstein

Deriving the Formula of Humanity (GMS, 427–437)

In *Groundwork* II, Kant tries to establish that if there is a supreme principle of morality, then it is (or is equivalent to) the Formula of Humanity. He offers a "derivation" of the principle: "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" (GMS, 429, emphasis omitted).

In broad outline, Kant's derivation (GMS, 427ff.) is not hard to discern. It takes shape against the background of his fundamental tenet that the supreme principle of morality would have to be a categorical imperative, that is, a principle binding on all of us no matter what our particular inclinations might be. First, Kant contends that if there is a supreme principle of morality (and thus a categorical imperative), then there is an objective end: something that is unconditionally valuable. A categorical imperative requires an unconditionally valuable "ground." Second, Kant claims that this unconditionally good thing would have to be humanity. In his view, therefore, if there is a supreme principle of morality, then humanity is unconditionally good. For Kant "humanity" does not refer to the class of human beings, but rather to a set of capacities. In the *Metaphysics of Morals*, Kant tells us that "the capacity to set oneself an end – any end whatsoever – is what characterizes humanity (as distinguished from animality)" (MdST, 392). So at the very least, if a being has humanity, then it has the capacity to set ends. Kant, it seems, uses "humanity" interchangeably with "rational nature" (see, for example, GMS, 439). In doing so he suggests that having humanity involves having certain rational capacities. Among them are the capacity to act on maxims and hypothetical imperatives, as well the capacity to act autonomously, that is, (roughly) to conform to self-given moral imperatives purely out of respect for these imperatives.[1] According to the third main claim in the derivation, if humanity is unconditionally good, then we must always treat it not merely as a means but also as an end. Therefore, if there is a supreme principle of morality, then we ought to do just what the Formula of Humanity says. So the supreme principle of morality, assuming there is one, must be this formula, or at least something equivalent to it.

This paper focuses on the derivation's first two steps, especially the second. Assuming that a categorical imperative would require there to be something unconditionally good, why must this unconditionally good thing be humanity, rather than something else? As Part II below points out, the argument Kant appears to invoke in response to this question, namely an argument by elimination (GMS, 428), is very disappointing. But Christine Korsgaard (1996, 106–132) and, later, Allen Wood (1999, 124–132) have claimed that this argument does not represent Kant's best effort at a derivation. Kant, they contend, actually undertakes a "regressive" argument. I concentrate on Wood's more recent account of the argument. According to a crucial part of Wood's account, Kant claims that whenever an agent sets an end, he must ascribe objective goodness to it.[2] But in ascribing objective goodness to his ends, Kant contends, the agent commits himself (rationally speaking) to holding his rational nature (that is, his humanity) itself to be unconditionally good. So, in sum, since a rational agent exercises his rational nature in setting ends, he must conclude that it is unconditionally valuable.

Part III questions whether Kant actually makes this argument. Contrary to Wood, I do not, for example, believe that, according to Kant's *Groundwork* position, whenever an agent sets an end, he is rationally compelled to ascribe objective goodness to it. I do not deny that it was open to Kant to appeal to the regressive argument or that

[1] Here I am following Thomas Hill, Jr. (1992, 38–41). Allen Wood presents a slightly different account of what Kant means by humanity (1999, 118ff.).

[2] As Allen Wood has pointed out to me, it is not obvious that Korsgaard attributes this claim to Kant. But I take as some indication that she does her statement that "in the argument for the Formula of Humanity, as I understand it, Kant uses the premise that when we act we take ourselves to be acting reasonably and so we suppose that our end is, in his sense, objectively good" (1996, 116). In any case, it is worth noting that Korsgaard's account of the regressive argument incorporates a robust notion of an objectively good end. On her account, Kant tries to show that if an agent takes himself to have objectively good ends, then he is committed to the unconditional goodness of humanity. But in order for an end to be objectively good in Kant's sense, she claims, it must be the object of rational choice, provide "reasons for action that apply to every rational being" (1996, 115) and be fully justified. All three criteria are apparent in 1996, 114–116.

the regressive argument is philosophically interesting. It is just not, I think, an argument that Kant himself unfurls.

Part IV tries to bolster Kant's argument by elimination with the help of material that he presents later in *Groundwork* II. Kant points out that rational nature is that which can possess a good will (GMS, 437). Since it is, he suggests, rational nature alone can be the unconditionally good "ground" of a categorical imperative. For all other candidates are such that if they had this status, the good will would be "subordinate" to them. But, claims Kant, "such a will cannot without contradiction be subordinated to any other object" (GMS, 437). Suitably interpreted, this reasoning strengthens Kant's argument by elimination, I contend. The argument turns out to be stronger than an isolated reading of GMS, 427–429, makes it appear. Of course, even if this claim is correct, Kant's argument is open to challenge. Part V focuses briefly on one objection to it.

Before exploring Kant's defense of the view that humanity must be the unconditionally good ground of any categorical imperative, we need to examine why he thinks a categorical imperative requires such a ground in the first place. Why, if there is a categorical imperative, must there be anything unconditionally good?

I.

In his initial step in the derivation of the Formula of Humanity, Kant claims that if there is a supreme principle of morality (and thus a categorical imperative), then there is something of absolute (i.e., unconditional) worth. In something of absolute worth alone "would lie the ground of a possible categorical imperative," and "if all worth were conditional and therefore contingent, then no supreme practical principle for reason could be found anywhere" (GMS, 428). But why could not a principle be unconditionally binding on us if nothing was unconditionally good?

According to Kant, an agent sets himself to do something, that is, determines his will, on the basis of his idea that doing this thing will enable him to secure some end. In Kant's view, all acting has an end (KpV, 34). Kant distinguishes between subjective and objective ends. Objective ends, if there are any, would hold for all rational beings. The idea of securing them would make available to all rational beings a sufficient ground (motive) for acting. But subjective ends do not give all rational beings grounds for securing them. These ends are such

that their "mere relation to a specially constituted capacity of desire on the part of the subject gives them their worth" (GMS, 428). Suppose a particular object is a subjective end. If an agent does not value this object, either in itself or as a means to something else, then it has no worth to him. And if the object has no worth to him, intimates Kant, then he does not have a ground to secure it. For him, it is not an end. Apparently, Kant has the following view: an agent has a sufficient ground to secure an object only if he values it – or at least is rationally compelled to value it. In the latter case, the agent is presumably able, through rational reflection, to come to value the object, thereby gaining a sufficient ground to secure it.

Against the background of this view, we can reconstruct the basis of Kant's claim that if there is a categorical imperative, then there must be an objective end – something absolutely valuable. A categorical imperative would be necessarily binding on all rational agents. But a principle could not be necessarily binding on all rational agents unless each of them necessarily had a sufficient ground (motive) at his/her disposal for obeying it. Take us, human rational agents. To say that a principle is binding on us is to say that we ought to (i.e., have an obligation to) conform to it. Kant, of course, holds that if an agent ought to do something, then she must be able to do it (e.g., KpV, 125; 159). But if an agent did not have a sufficient ground available to her for conforming to a rule, then she might not be able to conform to it. Thus, if not all rational agents necessarily have a ground for obeying a principle, then it cannot be a categorical imperative. As we noted, to have a ground for doing something an agent must, according to Kant, hold (or be rationally compelled to hold) the action or its effects to be valuable. Therefore, Kant seems to conclude, if there is a categorical imperative, then there must be something that everyone holds (or must hold) to be valuable: an objective end. There must be something that everyone, in every context, is rationally committed to valuing: something that is absolutely valuable or, equivalently, unconditionally good.[3]

II.

In the first step of his argument, Kant tries to show that if there is a categorical imperative, then there must be some object (or objects) that all rational agents must hold to be unconditionally good. In the sec-

[3] I am not convinced that this argument is successful. For detailed criticism of it, see Kerstein (2002, 47–54).

ond step, Kant claims that this something is humanity. Kant appears to pack his defense of this claim into the paragraph which begins: "Now I say that the human being and in general every rational being *exists* as an end in itself" (GMS, 428). For at this paragraph's end he concludes that if there is a "supreme practical principle for reason," then rational beings are of absolute worth (unconditionally valuable).

On its face, the defense seems inadequate. It appears to amount to an argument by elimination. Kant quickly dismisses three candidates for unconditional goodness: objects of inclinations, inclinations themselves, and beings "the existence of which rests not on our will but on nature," (GMS, 428) such as animals. Then he embraces humanity as alone suited to be an end in itself, that is, the unconditionally valuable ground of a categorical imperative.

His dismissal of rival candidates for unconditional goodness seems precipitate. For example, Kant says simply that "all objects of the inclinations have only a conditional worth; for if there were not inclinations and the needs based on them, their object would be without worth" (GMS, 428). But this statement requires defense that Kant does not here seem to offer. For an opponent might reasonably object that, regarding some such objects, Kant has things backwards: it is not the case that they are valuable (just) because we desire them, but rather the case that we desire them at least partly because they are (in themselves) valuable.[4]

Even if the remarks Kant here makes did eliminate these candidates for absolute goodness, the question would arise as to whether he is entitled to conclude that it is humanity he is looking for. After all, might not Kant have overlooked some other candidate for absolute goodness? What about the state of affairs of all rational agents being happy? How does Kant dismiss this possibility? This candidate is not itself an inclination. It need not be considered an object of an inclination. (We can easily envisage a world in which no one desires everyone – including his enemies – to be happy.) And everyone's happiness is not obviously something the existence of which would rest on nature rather than on our will.

Kant's argument that only humanity could be the absolutely valuable thing needed if there is to be a categorical imperative (supreme principle of morality) appears to suffer from serious shortcomings. His arguments against the other candidates he considers seem far too quick, and he does not explicitly consider at least one candidate that

[4] For a defense of this sort of objection, see Gaut (1997, 176 f.).

springs quickly to mind, namely everyone's being happy. On the face of it, Kant's argument by elimination does not seem very promising.

III.

Although he acknowledges shortcomings in Kant's argument by elimination, Wood does not despair of Kant's establishing that humanity must be the ground of any categorical imperative (1999, 124). For, according to Wood (1999, 125), Kant offers a positive argument for this claim in the very next paragraph.

A key step in Kant's positive argument, suggests Wood (1999, 127), is an attempt to establish the following: each rational agent must (is rationally compelled to) agree that in setting himself ends, he commits himself to the view that his rational nature is unconditionally valuable. Wood suggests that Kant's defense of this key step involves a "regress on conditions." Inseparable from an agent's setting ends for himself is his holding that they have a certain value. But an agent can, rationally speaking, place this value on his ends only on condition that he hold his own rational nature to be unconditionally valuable.

When we look more closely, Wood suggests, we find that this key step in Kant's argument rests on three claims. The first has to do with the kind of value the agent is committed to placing on the ends he sets. Whenever an agent sets himself an end, goes the claim, he must ascribe "objective" goodness to it. The second claim is that the agent must hold the source of the objective goodness of his ends to be his setting of them. What makes his ends good, he is rationally compelled to affirm, is that they are the objects of the exercise of a certain capacity, namely his rational choice.[5] Finally, if the source of the goodness of objectively good ends is his rational choice (rational nature), then the agent must hold his rational nature itself to be unconditionally good.[6] So, in sum, since a rational agent exercises his rational nature

[5] That Wood attributes the first two claims to Kant is manifest in the following passage: "Thus Kant's argument is based on the idea that to set an end is to attribute objective goodness to it and that we can regard this goodness as originating only in the fact that we have set those ends according to reason. The thought is . . . that rational choice of ends is the act through which *objective* goodness enters the world" (1999, 129). See also (1999, 127 and 129 f.).

[6] Wood makes clear that he attributes this claim to Kant when he states that "Kant does infer from the premise that rational nature is the source or ground of the objective goodness of all ends to the conclusion that rational nature itself is the underivative objective good, an end in itself" (1999, 130). See also (1999, 127 and 129).

in setting ends, he must conclude that it is unconditionally valuable. Even if we embrace this conclusion, we might wonder how to get from it to the further point that a rational agent must conclude that *everyone's* rational nature is unconditionally good.[7] We also might wonder how the *exclusivity* of rational nature as a possible ground for a categorical imperative gets established. That humanity is unconditionally valuable does not in itself seem to entail that nothing else is. But these are not issues for us to pursue here.

Also absent from our agenda is evaluating the philosophical plausibility of the argument that Wood (and Korsgaard before him) attribute to Kant. I find the argument (in both Wood and Korsgaard's versions) to be engaging and important, but as I have argued in detail elsewhere, I think it has serious shortcomings.[8] Here I concentrate on the question of whether there are adequate textual grounds for attributing the argument to Kant in the first place. The regressive argument may be Kantian in spirit. But I doubt whether he actually employs it in the *Groundwork*.

To begin, let us consider the first main claim that, according to Wood, Kant makes in defending the view that humanity is unconditionally valuable. For Kant, it seems, all cases of an agent's setting an end are cases of his exercising his capacity of rational choice (humanity). Kant implies that having humanity involves having the capacity to set ends.[9] The first claim is that *whenever* an agent sets an end, he must (is rationally compelled to) ascribe objective goodness to it. He must, says Wood, "regard it as something of value universally for all rational beings" (1999, 127). Let us assume that in Kant's view, when an agent sets an end, he must take it to be good in some sense. Why should we agree that in the *Groundwork* Kant embraces the view that he must take it to be *objectively* good? According to Wood, at GMS, 412–414, Kant "maintains that goodness, whether moral or nonmoral, is that which reason represents as practically necessary, and hence as an object of volition for all rational beings" (1999, 127). Something is an object of volition for all rational beings just in case all such beings must regard it as valuable, Wood implies. So, he suggests, at GMS, 412–414, Kant commits himself to the view that if an agent sets himself an end, whether she views the end to be good morally or non-morally, she is rationally compelled to hold it to be objectively good.

[7] Korsgaard (1996, 123) and Wood (1999, 131) each discuss this point.
[8] Cf. Kerstein (2002, 46–72).
[9] Cf. MdST, 392.

I do not see where in the specified pages Kant commits himself to this view.[10] Among the central claims Kant there makes are the following: All actions prescribed by an imperative (command of reason) are in some sense practically necessary, that is, good. Actions prescribed by a hypothetical imperative are good as means to some end; actions prescribed by a categorical imperative are good in themselves.[11] But it does not follow *from these claims* that if an agent sets herself an end, she is rationally compelled to hold it to be valuable for all rational beings. Suppose an agent sets herself an end that, she realizes, is an action commanded by a hypothetical imperative, but not a categorical imperative. It is a necessary means to some further goal she has. She would then have to agree that everyone is rationally compelled to hold the following: if someone, including the agent herself, has this further goal, then, as a means to this goal, it is good *for that person* to realize the end. But the agent would not be rationally compelled to hold that this end is something of value for each and every rational agent.

Let me illustrate this rather abstract point with an example. Suppose Sally sets herself the end of traveling to New York. This action, she realizes, is commanded by a hypothetical imperative, but not by a categorical imperative. Traveling to New York is a necessary means for visiting the Empire State Building, which is something she has had an inclination to do for a long time. (The hypothetical imperative at issue would be something like this: "If you want to visit the Empire State Building, then you ought to travel to New York.") If we grant the claims set out above, Sally would have to agree that everyone, including her, must hold the following: if a person has the end of visiting the Empire State Building, then it is good as a means to that end for that person to go to New York. But Sally would not be rationally compelled to agree that the end of traveling to New York is something that all rational agents must value.

In GMS, 412–414, Wood suggests, Kant embraces the view that if an agent sets himself an end he is rationally compelled to hold it to be objectively good. Some central claims Kant makes there do not entail this view. If there is a passage in GMS, 412–414, that seems to license this conclusion, it is perhaps the following: "Practical *good*,

[10] Thomas Hill, Jr. argues (2002, 244–274) that Kant does not embrace the view that in setting an end an agent always commits himself to the end's objective goodness. I focus more than does Hill on the issue of whether specific texts in the *Groundwork* provide evidence that Kant adopts the view in question.
[11] In this context Kant is using "imperative" as a success term; he is *conceiving* of imperatives as valid.

however, is that which determines the will by means of representations of reason, hence not from subjective causes, but objectively, i.e., from grounds that are valid for every rational being as such" (GMS, 413). If we identify "practical good" with any end an agent sets, then we seem to have evidence that Kant maintained the view in question. For then Kant appears to be suggesting that what determines an agent to pursue any end of his is a ground that is "valid for every rational being as such." And if what determines an agent to pursue any end of his is a ground valid for every rational being as such, then it seems that an agent is rationally compelled to view any end of his as valuable to every rational being.

But it is questionable whether in the passage cited from GMS, 413, Kant identifies "practical good" with any (and every) end an agent sets. There is no straightforward indication that he does. He uses the term end [*Zweck*] not once in the paragraph from which the passage is taken. Moreover, there is another interpretation of the passage available. By "practical good," I believe, Kant means something like "objective (practical) principles." In the two sentences preceding the passage, Kant is describing imperatives, which are objective practical principles expressed with the help of an "ought," since they are addressed to agents, such as human beings, who might fail to act in accordance with them. In the passage itself, Kant seems to be maintaining the following: if an agent does something at least partly on the grounds that an imperative commands it, then the agent is acting from grounds that are valid for every rational being as such. And that one maintains this does not require him to embrace the quite distinct view that in setting an end, an agent must take it to be objectively good. Suppose, for example, that Sally visits New York partly because the hypothetical imperative "If you want to visit the Empire State Building, then you ought to travel to New York," commands this action. (I describe her as acting "partly" from the imperative because a necessary ingredient of her acting from it at all is that she have the desire to visit the Empire State Building.[12]) Sally is acting from grounds that are valid for every rational being as such. For assuming that the principle in quotations is really an imperative, it is the expression of an objective principle–one valid for every rational being as such. The imperative is valid for us (human beings) in that if any of us have the end of visiting the Empire State Building, then, other things being equal, we ought to travel to

[12] At GMS, 428, Kant calls relative ends, such as Sally's end to visit the Empire State Building, grounds of hypothetical imperatives.

New York. Nothing here implies that in setting herself the end of visiting New York, Sally must embrace the view that the end is valuable to all rational agents. The interpretation on the table of GMS, 413, harmonizes well with what follows. In the very next sentence Kant suggests that practical good is "distinguished from the *agreeable*, as that which has influence on the will only by means of sensation from merely subjective causes, those which are valid only for the senses of this or that one, and not as a principle of reason, which is valid for everyone" (GMS, 413). An objective practical principle is, of course, a "a principle of reason, which is valid for everyone." It seems a stretch to claim that in *Groundwork*, 412–414, Kant endorses the view that in setting an end, an agent must hold it to be objectively good.

Actually, near the beginning of the derivation of the Formula of Humanity there is evidence that he does not endorse this view. Kant says:

> "The ends that a rational being proposes as *effects* of its action at its discretion (material ends) are all only relative; for only their relation to a particular kind of capacity of desire of the subject gives them their worth, which therefore can provide no necessary principles valid universally for all rational beings and hence valid for every volition, i.e., practical laws." (GMS, 427f.)

Let us suppose that a particular agent, John, correctly believes these claims to be true. John has a "material end," namely that of his attending the Super Bowl. According to Kant, it is *only* the relation of his end to the particular kind of capacity of desire he has that gives it its worth. In other words, it is only his setting it as an object to attend the next Super Bowl that gives his attending it any value. If he believes this, then why would he believe that everyone is rationally compelled to view his end as valuable? John would, it seems, be free to reason as follows: "Those who do not make it their object that I attend the next Super Bowl (and there are surely many) might be rationally compelled to acknowledge that my attending it is valuable to me. Yet it is not the case that they must take my attending it to be valuable to them." Moreover, Kant says that material ends "provide no necessary principles valid universally for all rational beings." But if material ends were, rationally speaking, objectively valuable, then they *would* provide at least one principle valid universally for all rational beings, namely the principle: "A material end is never to be treated as an object of no value whatsoever." This principle would have some practical bite. It would, for example, forbid an agent from failing to

promote John's end in cases in which doing so would not interfere with any of the agent's morally required action or pursuit of his own material ends. In the passage immediately preceding his derivation of the Formula of Humanity, not only does Kant fail to embrace the view that in setting an end, an agent must hold it to be objectively good, he implies that he rejects it.

The text of the *Groundwork* fails, I think, to imply that Kant embraces the first claim in the regressive argument for humanity as the ground of a categorical imperative. Kant does not, I contend, adopt the view that whenever an agent sets himself an end he must ascribe objective goodness to it.[13] If this contention is correct, then it is unsurprising that we find little evidence that he makes the second or third claims. Where does Kant actually make the argument that an agent must hold that the source of the objective goodness of his ends is his exercising his capacity of rational choice (rational nature) in setting them (claim 2)? Where does he actually make the argument that if an agent must hold this, then he must also hold his rational nature itself to be unconditionally good (claim 3)? According to Wood, Kant makes these arguments in order to defend his contention at GMS, 429, that each of us necessarily represents his existence as that of an end in itself, that is, as that of something that is unconditionally valuable. I believe that it was open to Kant to make the arguments. But I do not find compelling textual evidence that he actually does.

Indeed, Kant's claim that each of us necessarily represents his existence as that of an end in itself, as well as the surrounding text, admits of an interpretation that does not invoke the regressive argument at all.[14] In the paragraph at the end of which Kant sets out the Formula of Humanity, he says:

> "The ground of [the moral principle] is: *Rational nature exists as end in itself*: The human being necessarily represents his own existence in this way; thus to that extent it is a *subjective* principle of human actions. But every other rational being also represents his existence in this way consequent on just the same rational ground that also holds for me;* thus it is at the same time an *objective* principle from which, as a supreme practical ground, it must be possible to derive all laws of the will." (GMS, 429)

[13] Both Korsgaard (1996, 115) and Wood (1999, 128f.) appeal to passages outside of the *Groundwork* for support of their view that, according to Kant, if an agent sets an end, he holds it to be objectively valuable. Both invoke, for example, Kant's discussion of goodness and well-being in *The Critique of Practical Reason* (KpV, 58–61). For an interpretation of this discussion according to which Kant is *not* there committing himself to the view in question, see Hill (2002, 262ff.).

[14] This contention challenges Korsgaard as well as Wood. See Korsgaard (1996, 122f.).

The note indicated in the second sentence reads as follows: "Here I put forward this proposition as a postulate. The grounds for it will be found in the last Section" (GMS, 429). Kant's claim that the human being necessarily represents his own existence as an end in itself might be a claim about human nature. Kant might be suggesting that it is natural for human beings to think of themselves as superior to non-rational beings, including other animals.[15] But perhaps Kant wants to leave it open that non-human rational agents might not, as a rule, act against the background of such a view. That would explain why Kant calls doing so a "subjective" principle of acting. In any case, the plausibility of the claim that human beings do indeed represent themselves as superior to non-rational beings would not seem to depend on an argument regarding the ultimate source of value in the world.

But what are we then to make of the rest of the passage, in particular of Kant's remark that "every other rational being also represents his existence in this way consequent on just the same rational ground that also holds for me"? This remark seems to me to contain two distinct claims. The first is that we, human rational agents, have a rational ground, that is, are rationally compelled, to represent all rational agents as unconditionally valuable. The second is that non-human rational agents are also rationally compelled to view all rational agents as having such a value. But to this point in the *Groundwork*, Kant has proven neither of these claims. By his own lights, he has shown merely that *if* one assumes there to be a categorical imperative (supreme principle of morality), then one must hold humanity to be an end in itself. He has not shown (nor tried to show) that anyone who does not make this assumption is rationally compelled to hold humanity to be an end in itself. In other words, Kant has not proven the validity of the categorical imperative (Formula of Humanity). That is a task that he puts off until *Groundwork* III. There he tries to show that all agents (rational beings with a will), not merely human agents, must take themselves to be free and thus to be bound by the moral law.[16] It is therefore not at all surprising that, in a note, Kant calls the remark in question a "postulate" and suggests that he will defend it in *Groundwork* III.

Our reflections thus far have not left us with a good impression of Kant's *Groundwork* derivation of the Formula of Humanity. He seems to offer a very weak argument by elimination in defense of his claim that the unconditionally good ground of a categorical impera-

[15] See Anth, 127.

[16] See GMS, 447f.

tive would have to be humanity. If I am correct, he simply does not present the more philosophically interesting regressive argument attributed to him by Wood. However, things might be brighter than they appear. Well after he unfolds his argument by elimination (but still in *Groundwork* II), Kant sets out considerations that supplement and reinforce it. In particular he appeals to the notion of a good will in arguing that if there is a supreme principle of morality, then humanity, but not anything else, is an end in itself. Reflection on this appeal will help us to see that Kant's argument by elimination is stronger and more interesting than it might seem to be.

IV.

At GMS, 437, Kant begins a summary of a main line of argument he has developed up to this point in the *Groundwork*. "Now we can end," he announces, "at the place from which we set out at the beginning, namely with the concept of an unconditionally good will." In the next paragraph, he says the following:

> "Rational nature is distinguished from the rest of nature by this, that it sets itself an end. This end would be the matter of every good will [. . .] Now, this end can be nothing other than the subject of all possible ends itself, because this subject is also the subject of a possible absolutely good will; for, such a will cannot without contradiction be subordinated to any other object." (GMS, 437)

At GMS, 428, let us recall, Kant seems to dismiss precipitately three candidates for the unconditionally good ground of a categorical imperative: inclinations themselves, the objects of inclinations, and beings "the existence of which rests not on our will but on nature." In the passage just cited, Kant suggests an argument that might bolster his elimination of these candidates. The argument takes shape against the background of Kant's view of the value that "common rational moral cognition" attributes to a good will. In our ordinary moral thinking, claims Kant, we hold that such a will is unconditionally good. An impartial rational spectator would, we believe, judge it to be good in every possible context in which it exists.[17] Moreover, we believe that a good will is preeminently good.[18] That we believe this presumably implies that, in our view, nothing that is devoid of a

[17] See GMS, 393.

[18] See GMS, 394 and 401.

good will is as good as something that has such a will. The argument unfolds as follows. Whatever constitutes the ground of a categorical imperative must be compatible with the value we attribute to a good will (i.e., unconditional and preeminent goodness). But let us suppose that any of the rivals to humanity were the ground of a categorical imperative. We would have to acknowledge that a good will would (in some circumstances) be "subordinated" to the rival and that a good will would thereby fail to have the value we attribute to it. We would land in contradiction. Therefore, we have license to dismiss any of the rivals as possible grounds for a categorical imperative. As is perhaps already evident, the material that is to supplement Kant's argument by elimination is material Kant develops before his presents the argument. I will not appeal to any positions Kant develops after he does so—for example his view that humanity has dignity, that is, unconditional and incomparable value.

Filling in the supplemented argument's details requires elaboration of the concept of a good will as well as of the notion of one thing's being subordinated to another. The latter task is in a sense easier, since Kant simply does not tell us precisely what this notion amounts to. But I think it reasonable to assume that he would endorse the following claim: If x is subordinated to y, then x is less valuable than y and we thereby have sufficient grounds to use x in whatever way is necessary in order to maintain y. So for example, if plants are subordinated to rational beings, then they are less valuable than rational beings, and we have sufficient reason to harvest the former in order to preserve the latter.

Regarding a good will, it suffices for our purposes to take note of two ways in which Kant seems to employ the notion as it applies to us, agents who can indulge their inclinations and thereby act contrary to the moral law. 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" (GMS, 411), presumably of actions done from duty, which he has previously stated to have "unconditional and moral worth" (GMS, 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.[19] According to

[19] This reading of "good will" would have to be broadened to accommodate Kant's view that perfectly rational beings such as God cannot act from duty. To them the "ought" of duty does not apply, since their willing is necessarily in accord with the law. See GMS, 414. We might attribute to Kant the view that these beings have a

a second usage of "good will," it 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 on this usage just in case she is committed to doing what duty requires, not just in this or that particular action, but overall. Presumably if an agent has this commitment, then she will 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.) Kant intimates that having a good will amounts to having a certain kind of character in the first paragraph of *Groundwork* I. 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 which is to make use of these gifts of nature, and whose distinctive constitution is therefore called *character*, is not good" (GMS, 393).[20] Sometimes Kant employs what we might, following Karl Ameriks (1989, 54–59), call the "whole character" conception of a good will. I believe Kant to be employing the whole character conception of a good will in the passage we are discussing (GMS, 437). In any case, it is that conception of a good will that might help him to bolster his argument by elimination.

Let us now return to our (GMS, 437–based) supplement to Kant's argument by elimination. Suppose that beings "the existence of which rests not on our will but on nature," say, species of wild animals, were the ground of a categorical imperative, namely one commanding us never to eradicate currently existing species of such animals. We would be committed to the view that such beings were not only unconditionally but also preeminently valuable. Otherwise, we might sometimes lack sufficient grounds to abide by the principle not to eradicate them. And if we might lack such grounds, we cannot take the principle to be a categorical imperative.

We would presumably lack sufficient grounds to abide by the principle in question when doing so would conflict with maintaining some-

good will (engage in unconditionally good willing) just in case they act "for the sake of the law." Presumably such beings are capable of doing this. And Kant does not seem averse to the idea that acting from duty is a species of acting for the sake of the law.

[20] 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" (GMS, 398f.). 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.

thing that was also unconditionally valuable, but *more* valuable than a wild animal species. That something might, for example, be a person. Say that the only way to save someone's life was to kill the last two remaining representatives of a bird species in order to make a medicine for him. If we believe correctly that persons are more valuable than species of wild animals, then, in these circumstances, we do not have sufficient reason to act in accordance with a principle commanding us never to eradicate the latter. It seems plausible to grant the *possibility* that more than one kind of thing is unconditionally good as well as that one unconditionally good thing is better than another. After all, what basis do we have for denying it? Yet if we grant this possibility, we find that preeminent as well as unconditional goodness is necessary to ground a categorical imperative.

Getting back to our example, if species of wild animals can serve as the ground of a categorical imperative, then they must be unconditionally and preeminently good. Their being preeminently good implies that nothing that is not a species of wild animals is as good as something that is. But if species of wild animals have this value, then a good will's value is subordinate to their value. They are worth more than it is, and so we have sufficient reason to abandon it in order to preserve them. As Kant suggests it would at GMS, 437, this subordination of a good will to species of wild animals results in a contradiction. For we have been assuming that a good will is unconditionally and preeminently good. But a good will cannot both be less valuable than species of wild animals and, as the notion of preeminence implies here, more valuable than they are.

It would, I believe, be unproblematic to illustrate via an analogous chain of reasoning how Kant might eliminate other candidates for the ground of a categorical imperative, including inclinations and the objects of inclinations. Even a candidate Kant does not seem to consider in his argument by elimination, namely everyone's being happy, is vulnerable to this reasoning.

Suppose that everyone's being happy were the ground of a categorical imperative, namely one commanding us to maximize the aggregate welfare. We would then be committed to the view that everyone's being happy was both unconditionally and preeminently valuable. If it were merely unconditionally valuable, we might sometimes lack sufficient grounds (i.e., rational justification and thus motivation) to abide by the principle to maximize aggregate welfare. For example, we might lack sufficient grounds for abiding by this principle when doing so would prevent us from securing something more valuable

such as, perhaps, the existence of persons. And if we might be without adequate reason to conform to a principle, then we cannot take it to be a categorical imperative. But if everyone's being happy has both unconditional and preeminent value, then a good will's value is obviously subordinate to it. (For Kant, of course, not everyone who has a good will is happy and not everyone who is happy has a good will. See, for example, GMS, 442.) The object constituted by everyone's being happy is worth more than a good will, and so we have sufficient reason to destroy the latter in order to promote the former. But this conclusion forces us into a contradiction. We are assuming along with Kant that a good will is unconditionally and preeminently good. Yet a good will cannot both be less valuable than everyone's being happy and, as the notion of its preeminence implies here, more valuable than it is.

Of course, the supplemented argument by elimination just brought to bear against a version of utilitarianism is far from invulnerable to attack. It is obviously open to philosophers to disagree with Kant's view that, according to ordinary moral thinking, a good will is unconditionally and preeminently good. Moreover, it remains to be seen how the supplemented argument avoids dismissing Kant's own candidate for the ground of a categorical imperative. How is it that humanity itself does not get eliminated from contention?

Initially, it seems that it would. After all, if humanity is to be the ground of a categorical imperative, then we must hold it to be not only unconditionally good, but also preeminently good. Otherwise we leave open the possibility of there being other, better, unconditionally good things–things that we would, rationally speaking, have to preserve even at the cost of disobeying the principle grounded by the value of humanity. But if we hold humanity to be preeminently as well as unconditionally good, then we must, it seems, conclude that a good will is subordinate to it, contradicting our assumption that the good will is preeminently good.

Fortunately, this argument ignores the special relationship that obtains between humanity and a good will. Every being who has a good will necessarily has humanity. For having a good will entails having capacities constitutive of humanity, for example, the capacity to act on categorical imperatives. An agent cannot have an overall commitment to do what duty requires according to Kant unless she has the capacity to act on principles that specify what it requires, namely moral rules. Much as it would be impossible to be an excellent classical pianist without being a pianist, so it would be impossible to have a good will without having humanity. With this point in view, we are able to see

that the supplemented argument by elimination does not, in the way the objection alleges, throw the baby (humanity) out with the bath water (rival candidates for the ground of a categorical imperative). If we hold humanity to be preeminently good, then we believe that nothing that is devoid of humanity is as good as something that has it. A good will, of course, is not devoid of humanity. So holding humanity to be preeminently good does not force us to conclude that a good will is subordinate to it, that is, has less value than it does.

Here it makes sense to object that, even if this last point is correct, humanity would get eliminated as a candidate for the ground of a categorical imperative. For the idea that humanity is unconditionally and preeminently good contradicts the notion that a good will is unconditionally and preeminently good. Two different things cannot both be preeminently valuable.

Given the notion of preeminence that (at least on my reading) Kant employs, it turns out that two different things can both be preeminently good. Humanity and a good will are indeed two different things; for one can have humanity without having a good will. Not every being possessed of rational nature (humanity) is committed overall to doing what moral principle requires. Some of us do not have excellent character. But there is no contradiction in maintaining that humanity is preeminently good and all the while holding that a good will is. According to the former claim, nothing that is devoid of humanity is as good as something that has it. According to the latter, nothing that is devoid of a good will is as good as something that has it. The latter claim implies that humanity without a good will is not as good as humanity with a good will. But there is no contradiction in maintaining both that nothing devoid of humanity is as good as something with it and, at the same time, that humanity devoid of a good will is not as good as humanity with one. In much the same way, it is not self-contradictory (though it may be false) to maintain both that no frying pan that is not copper is as good as any frying pan that is copper and, at the same time, to hold that every copper frying pan that is 3 mm thick is better than any copper frying pan that is less thick.

V.

Even when supplemented as I suggest, Kant's argument by elimination appears to be closer to Kant's intentions concerning the *Groundwork* derivation of the Formula of Humanity than the regressive argument

attributed to him by Korsgaard and Wood. For not only does Kant explicitly embrace the reasoning that bolsters the argument (namely at GMS, 437), but this reasoning itself relies only on concepts that Kant develops before the derivation is complete.

Of course, the supplemented argument from elimination and thus the derivation as a whole is, at best, only as convincing as Kant's view that a good will is unconditionally and preeminently valuable. This view is surely in need of defense.[21]

But let me conclude by considering a further challenge to the supplemented argument. If the reasoning in Part IV is sound, the argument avoids two pitfalls. It does not descend into self-contradiction; for there is nothing inconsistent in maintaining as the argument does that both humanity and a good will are unconditionally and preeminently good. Moreover, the argument does not imply that a good will is subordinated to humanity. So it avoids thereby eliminating humanity as a possible ground for a categorical imperative.

However, a serious question remains: What justification does Kant have for holding that it is beings with rational nature who constitute the ground of a categorical imperative, rather than maintaining that it is merely beings with a good will who do so? What, for example, permits Kant to reject the view that it is not all of us, but rather only those of us with a good will, who never ought to be treated merely as means? Consider a principle that Kant does not discuss, namely what I call the Formula of the Good Will: "So act that you treat a good will, 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." It makes sense to stipulate that, according to this formula, having a good will necessarily involves having an overall commitment to doing what Kant's Formula of Universal Law requires. For Kant's derivation of the Formula of Universal Law takes place well before the argument by elimination. The supplemented argument by elimination seems to leave open the possibility that a good will serves as a ground for the Formula of the Good Will. So why must we say that it is humanity rather than a good will that remains viable as a ground for a categorical imperative?[22]

[21] But the supplemented argument does not depend on a further controversial view Kant holds (GMS, 393), namely that *only* a good will is good without qualification.

[22] In reply, one might focus attention on the fact that Kant's derivation of the Formula of Humanity occurs after his derivation of the Formula of Universal Law. One might then argue that Kant implicitly holds the following: whatever is to serve as the ground of a categorical imperative must serve as the ground of a principle equivalent to the Formula of Universal Law, that is, a principle that requires or permits just those actions that are required or permitted by the Formula of Universal Law. But the good will would serve as the ground of a principle, namely the Formula of the Good Will, that is not equivalent to the Formula of Universal Law. Therefore, concludes the argument, the good will is not the ground of a categorical imperative. Perhaps Kant would embrace this argument, but I do not find it philosophically promising. Kant, of course, suggests that the Formula of Humanity is equivalent to the Formula of Universal Law (GMS, 436). But I know of no substantive interpretation of the two principles according to which they turn out to be equivalent. So I fear that this sort of argument would likely not only eliminate the good will as a possible ground of a categorical imperative, but humanity as well.

Our response to this objection depends on our interpretation of a thorny issue concerning a good will. According to Kant, an agent can never be sure that he (or anyone else) has a good will.[23] A commitment to morality over the fulfillment of inclinations is constitutive of a good will; but introspection cannot reveal the presence of any such commitment, he implies. Even if one has never violated his duty, one might nevertheless lack a good will. For one's conformity to duty might result from a fortunate harmony in his case between the dictates of prudence and the requirements of morality, rather than from a genuine commitment to the moral law.[24] But can we ever be sure that an agent *lacks* a good will? The answer to this question is not obvious. What if, for example, Sue has manifested, and continues to this moment to exhibit, a pattern of breaking promises, apparently just for her own financial gain? In Kant's view the Formula of Universal Law forbids acting on maxims of false promising for financial gain (GMS, 422). Moreover, Kant assumes that it is a simple matter for all of us to determine what morality demands.[25] So can we rest assured that Sue (or anyone else) does not currently have a good will?

This is not the occasion to try to answer this question. But much depends on which answer turns out to be correct. Suppose that we *can* be sure that an individual, say Sue, now lacks a good will. In this case, the Formula of Humanity and the Formula of the Good Will would not, for practical purposes, be equivalent. The former but not the latter would forbid treating Sue merely as a means.[26] Humanity

[23] See GMS, 407f.

[24] See RGV, 36f.

[25] See, for example, GMS, 404, and KpV, 36. This view seems to me to be unpersuasive, as I explain in Kerstein (2002, 119–129).

[26] At this point, it might be tempting to make the following argument. "Even if we are sure that Sue does not have a good will now, it is open to her to develop one in the future. On the basis of this potential she has, the Formula of the Good Will would forbid us from treating her merely as a means." But this argument misses the mark. For the Formula of the Good Will does not forbid us from treating beings with the potential for having a good will merely as means, but rather from treating beings with a good will merely as means. And here we are assuming that we know Sue not to have a good will.

and a good will would constitute competing grounds for categorical imperatives, and the supplemented argument by elimination would be incomplete.

Actually, if we can know that an individual lacks a good will, then the argument would be worse than incomplete. It would discount humanity as a possible ground for a categorical imperative, namely for the Formula of Humanity itself. It is easy to imagine situations in which this principle and the Formula of the Good Will would deliver incompatible moral verdicts. As we just noted, the latter would imply that treating Sue merely as a means is morally permissible while the former would entail that it is not. But in such a case an agent would find herself without sufficient grounds to abide by the Formula of Humanity. Why should she privilege the dictates of the Formula of Humanity over those of the Formula of the Good Will? It is true that the Formula of Humanity is (supposedly) grounded on something unconditionally and preeminently valuable, but the Formula of the Good Will is grounded on something that has this value and is (supposedly) even better.[27] Of course, if there is a possible situation in which an agent would lack sufficient grounds to abide by a principle, then this principle is not a viable candidate for a categorical imperative.

But suppose that we *cannot* be sure that any particular person currently lacks a good will. In order to abide by the Formula of the Good Will, it seems, we would then treat as ends in themselves all beings with humanity. For all such beings *might* have a good will. Humanity, says Kant, is "the subject of a possible absolutely good will" (GMS, 437). In order to *insure* our compliance with the Formula of the Good Will, we would have, therefore, to treat everyone with humanity as if he/she had a good will. So in reply to the objection, one might say that the Formula of the Good Will is for practical purposes equivalent to the Formula of Humanity. Since it is, a good will and humanity are not really competing grounds for a categorical imperative.

If Kant's considered view is that we can never know whether an individual lacks a good will, his supplemented argument by elimination avoids a significant obstacle.[28] In any case, this argument deserves further attention, I believe. It is not only philosophically interesting, but also well-grounded in Kant's text.

[27] See IV above.

[28] I would like to thank the other contributors to this volume for very helpful discussion of this paper.

Literature

Kant's writings

Kant's writings will be cited according to the pagination of *Kants gesammelte Schriften*, Akademie Ausgabe (Berlin: W. deGruyter, 1902-) (abbreviated as 'AA'). All English translations are based on *The Cambridge Edition of the Works of Immanuel Kant* (Cambridge University Press, 1992–).

Anth *Anthropologie in pragmatischer Hinsicht*, AA VII
GMS *Grundlegung zur Metaphysik der Sitten*, AA, IV
KpV *Kritik der praktischen Vernunft*, AA, V
MdST *Metaphysik der Sitten, Tugendlehre*, AA,VI
RGV *Die Religion innerhalb der Grenzen der blossen Vernunft*, AA, VI

Other works

Ameriks, Karl (1989): "Kant on the Good Will," in: Höffe, Otfried (Ed.): *Grundlegung zur Metaphysik der Sitten: Ein kooperativer Kommentar*, Frankfurt a.M., 45–65.

Gaut, Berys (1997): "The Structure of Practical Reason," in: Cullity, Garrett/ Gaut, Berys (Ed.): *Ethics and Practical Reason*, Oxford, 161–188.

Hill, Thomas, Jr. (1992): *Dignity and Practical Reason*, Ithaca: Cornell University Press.

Hill, Thomas, Jr. (2002): *Human Welfare and Moral Worth*, Oxford: Oxford University Press.

Kerstein, Samuel J. (2002): *Kant's Search for the Supreme Principle of Morality*, Cambridge: Cambridge University Press.

Korsgaard, Christine M. (1996): *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press.

Wood, Allen W. (1999): *Kant's Ethical Thought*, Cambridge: Cambridge University Press.