

# Mimicking Korsgaard

Jon Garthoff  
Northwestern University

## 1. Introduction<sup>1</sup>

1996 was a banner year for Christine Korsgaard. It saw the publication first of her book *The Sources of Normativity* and then of her collection of essays *Creating the Kingdom of Ends*.<sup>2</sup> The former was hailed as “strikingly original”<sup>3</sup> and as a “modern classic”<sup>4</sup> by reviews not atypical of its reception; and the latter has been described, in my view quite appropriately, as a “deeper and more wide-ranging presentation of the author’s views”.<sup>5</sup> Few philosophers have impacted the profession so strongly in a single year.

This impact has rarely manifested, however, as persuasion. The centerpiece of Korsgaard’s work—her thoroughgoing Kantian constructivism about value—has few wholehearted adherents.<sup>6</sup> And while a great many philosophers have been strongly influenced by Korsgaard’s work, there is no consensus that she has gotten something importantly and distinctively correct. I believe, however, that many of Korsgaard’s insights concerning the nature of value and its relationship to when actions are morally required *are* importantly and distinctively correct; and in this essay I attempt to elucidate and motivate those insights. I attempt this, moreover, within an account of value very different from Korsgaard’s, and very different from Kant’s as well.<sup>7</sup> I endorse non-naturalist realism, not constructivism, about value. I do not attempt to defend this realism in a systematic way here,<sup>8</sup> but I mention my commitment in order to expose how little one need accept Korsgaard’s overall framework in order to accept some of her core insights. Korsgaard’s work has a compelling structure that, like most outstanding work in philosophy, can be divorced from its least plausible components.

I focus my discussion, as several important recent commentaries have done, on what Korsgaard calls her “regress argument”.<sup>9</sup> I do not believe this argument succeeds in all of its aspirations; I reject

Korsgaard's conclusion that all value is conferred through rational willing, for example, and I deny her assertion that the regress argument has antiskeptical force. I thus endorse some previously articulated criticisms of the regress argument,<sup>10</sup> and I offer a new objection to it in Section 4 below. I maintain, however, that there is a successful analog of the argument, which a value realist can and should endorse. And if I am right that this argument succeeds where Korsgaard's regress argument fails, then—as I argue in Sections 5 and 6—we can recover a series of her claims by mimicking, within a realist value theory, her argumentative moves with respect to the nature of value and its relationship to moral requirements. I argue that we can recover, in particular, both her claim that the exercise of rational capacities helps sustain the final value of aims and her claim that all persons are required not to violate Kant's formula of humanity. But first I must introduce Korsgaard's regress argument; and to do that, I must explain the background in Kant's *Groundwork of the Metaphysics of Morals* that inspired it.

## **2. The Good Will and Humanity**

The central claim of Kant's theory of value is his statement at the opening of *Groundwork I* that a good will, and only a good will, is good without limitation or qualification:

It is impossible to think of anything in the world, or indeed even beyond it, that could be considered good without limitation except a good will.<sup>11</sup>

Call this claim the “good will thesis”. I do not here assess the good will thesis, and so I will not attempt here to present a sophisticated interpretation of Kant's argument for it. The crux of this argument is, however, quite simple: it consists in the observation that anything other than a good will can be used to achieve bad aims.<sup>12</sup> This is clearly true of paradigmatically instrumentally valuable things like money and power. More innovative and interesting is Kant's extension of this insight by qualifying the value of things, such as happiness and excellence, which we value not instrumentally but *finally*, or for their own sake. The judgment Kant trades on here is that which is at work when we think a sadistic torturer is mistaken to value the satisfaction he derives from inflicting pain, or when we think that it would be better if a

sadistic torturer were less resolute, since he would then less effectively pursue his sadistic aims. Kant's explanation of these judgments is that things we otherwise regard as valuable for their own sake—satisfaction in the first case, virtue in the latter—fail to be valuable when conjoined with a bad will. This entails that the value of these things is qualified. But Kant claims that there are no such considerations qualifying the value of a good will. Kant thus asserts that a good will is the only unqualified good on the ground that it is the only thing that cannot be used, as one might say in the idiom of comic books, for the “forces of evil”.

The central claim of Kant's theory of moral obligation is expressed by his famous 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.<sup>13</sup>

As Kant uses it, “humanity” is a term of art. The English word “humanity” typically refers to membership in the human species or to benevolent dispositions towards others. But for Kant, “humanity” refers to the rational capacities of embodied individuals.<sup>14</sup> The formula of humanity thus states a requirement about how these rational capacities must be treated. The formula of humanity explicitly limits the scope of moral requirements, moreover, to cases where humanity is instantiated in persons; it is in your own *person* or in the *person* of any other that humanity cannot be treated as a mere means. Thus for Kant only full-blown rationality—that is, personhood—generates moral obligations.<sup>15</sup>

Given this very brief account of Kant's theories of value and obligation, one may raise a somewhat naïve question: why are we obligated to treat all humanity as an end in itself, when only a good will is claimed to be good without qualification? Why not claim instead that only persons with a good will are worthy of our respect?

This question invites a search for a bridge principle to explain why humanity is itself worthy of respect, even though it does not always, or even typically—or possibly ever—manifest as a good will. One might appeal, for example, to Kant's claim that we can never be certain when a particular person, including ourselves, has a good will.<sup>16</sup> This uncertainty is a consequence of two facts about obligation as Kant understands it: that it is possible to act permissibly without being motivated by respect for humanity, and that neither our own motives nor

the motives of others are transparent to us. We might try, on Kant's behalf, to exploit this inscrutability of good willing in an effort to bridge the gap between the good will thesis and the formula of humanity. If we can never be sure which rational beings have a good will, the thought is, then the only way to be sure that we respect every person with a good will is to respect every person.<sup>17</sup>

It is a mistake, however, to try to bridge the gap in this way. For one thing, some rational beings manifestly lack a good will, since like the sadistic torturer they are known to have impermissible aims;<sup>18</sup> yet the formula of humanity entitles them to be treated with respect too. And perhaps more importantly, the formula of humanity would have the wrong modality if it were justified by our inauspicious epistemic situation. The point of complying with the formula of humanity is not to make sure that one does not ever fail to respect a good will; it is, rather, that all rational beings are worthy of our respect.

It may be tempting to appeal instead to a bridge principle that emphasizes rationality's potential to instantiate a good will.<sup>19</sup> All rational beings must be respected, on this suggestion, because they are all capable of instantiating a good will. Perhaps there is a defense for Kant's theory along these lines. I will not pursue the issue here, however, since I do not believe that the best defense of the formula of humanity proceeds from the putatively unqualified value of a good will. Here I note only that this strategy will be complicated. It is not in general the case that if *Ys* have a distinctive value and *Xs* are the only potential *Ys*, then *Xs* also have that distinctive value. Consider, as plausible counterexamples to this sweeping claim, the relationship between oil paintings and canvasses or that between adult humans and human embryos. If humanity must be respected because it is a potential good will, there must be a further feature of its relationship to the good will, beyond the mere fact that only rational capacities can instantiate a good will, that helps explain this fact.

Although I do not attempt to defend this exegetical claim here, my own view is that searching for principles to bridge the gap between the value of a good will and the value of humanity misconstrues the argument of the *Groundwork*.<sup>20</sup> The good will thesis appears earlier in the text, but Kant appears to offer an independent argument for the formula of humanity at G428-429, just before he explicitly states that principle. In this compact and somewhat obscure passage, Kant makes no mention of a good will or its value; and so it is plausible to think that

Kant's case for the formula of humanity does not rely on the good will thesis.

### 3. The Ambition of the Regress Argument

It is noteworthy, however, that Korsgaard's interpretation of the argument at G428-429—surely the most influential recent discussion in English—*does* appear to rely on the good will thesis. Versions of Korsgaard's account of her regress argument for the formula of humanity appear in a variety of locations in her work.<sup>21</sup> Before discussing the argument, I quote extensively from Korsgaard's papers, in an effort to convey the spirit of the argument as Korsgaard understands it. Here is Korsgaard, for example, in the opening chapter of *Creating the Kingdom of Ends*:

A thing has conditional value if its value depends on whether certain conditions hold. For instance, the value of the means depends on the value of the end it serves; and the value of an object of desire depends on whether satisfying the desire will really contribute to the person's happiness. Even happiness is not valuable in all cases, and so is conditional. A thing has unconditional value if it has its value in itself and so has it under all conditions. Ultimately all value must spring from a source which is unconditionally valuable, for as long as we can question the value of something, we have not reached the end of its conditions.<sup>22</sup>

In "Kant's Formula of Humanity":

[W]hat makes the object of your rational choice good is that it *is* the object of a rational choice. ... [Kant's] idea is that rational choice has what I will call a value-conferring status. ... We act as if our own choice were a sufficient condition of the goodness of its object. ... If you view yourself as having a value-conferring status in virtue of your power of rational choice, you must view anyone who has the power of rational choice as having, in virtue of that power, a value-conferring status. ... Thus, regressing upon the conditions, we find that the unconditioned condition of the goodness of anything is rational nature, or the power of rational choice. To play this role, however, rational nature must itself be something of unconditional value—an end in itself. This means, however, that you must treat rational nature wherever you find it (in your own person or in that of another) as an end. This in turn means that no choice

is rational which violates the status of rational nature as an end: rational nature becomes a limiting condition (G437-38) of the rationality of choice and action. It is an unconditional end, so you can never act against it without contradiction.<sup>23</sup> [Korsgaard's emphasis and citation.]

In "Aristotle and Kant on the Source of Value":

The objects of inclination have only a conditioned value, [Kant] says, for their worth depends on the inclinations themselves (the things we desire are good because we desire them, not the reverse). The inclinations, however, cannot confer value on their objects, for they are not themselves unconditionally valuable. ... The existence of an inclination is not enough to make its object good, for the inclination itself may be bad. ... The argument is in a simple sense transcendental: we regard some of our ends as good, even though they are obviously conditional; there must be a condition of their goodness, a source of their value; we regard them as good whenever they are chosen with full rational autonomy; so full rational autonomy itself is the source of their value. Since this holds for other rational beings as well as myself, I cannot act against their rational autonomy without violating my own; and so it turns out to be a good will that is the source of all value.<sup>24</sup>

And, finally, in "Two Distinctions in Goodness":

Instruments ... can only be conditionally valuable. If the conditions of their goodness are met, however, they can be good objectively. The more important point is about things valued as ends. These are also conditionally or extrinsically good. In particular, happiness, under which Kant thinks all our other private purposes are subsumed, is only conditionally good ... . But although happiness is conditionally valuable, it is, when the condition is met, objectively good. In order to see this it will help to keep in mind Kant's other uses of the unconditioned-conditioned distinction. If anything is conditioned in any way, reason seeks its condition, continually seeking the conditions of each condition until it reaches something unconditional. It is this characteristic activity of reason that generates the antinomies of theoretical speculative reason in the *Critique of Pure Reason*. The usual example is causal explanation – if we explain a thing in terms of its cause, we then go on to explain the cause itself in terms of a cause, and this process continues. Reason does not want to rest until reaches something that needs no explanation (although this turns out not to be available): say, something that is a first cause or is its own cause. A

causal explanation truly satisfying to reason would go all the way back to this evident first cause, thus *fully* explaining why the thing to be explained must be so. These are familiar moves in philosophy, so there is no need to belabor the point. To apply it here, it is only necessary to point out that just as to explain a thing fully we would have to find its unconditioned first cause, so to *justify* a thing fully (where justify is “show that it is objectively good”) we would have to show that all the conditions of its goodness were met, regressing on the conditions until we came to what is unconditioned. Since the good will is the only unconditionally good thing, this means that it must be the source and condition of all the goodness in the world; goodness, as it were, flows into the world from the good will, and there would be none without it.<sup>25</sup> [Korsgaard’s emphases.]

Later in this section I briefly explicate Korsgaard’s regress argument and articulate why I do not believe that it succeeds as she understands it. I do not, however, attempt to present a comprehensive discussion of the regress argument.<sup>26</sup>

My principal aim in what follows is instead to distill the elements of Korsgaard’s argument that I believe withstand objection and to distinguish these from the elements of the argument that I believe should be abandoned. Because this is my aim, I couch the claims in my reconstruction of Korsgaard’s argument in a way that is neutral between her constructivism and the realism that I advocate. The idea is to present an argument true to Korsgaard’s intent—one formulated such that she or another Kantian constructivist could endorse it—but that admits of an alternative, realist interpretation. My reconstruction of her argument is thus an archetype, for which her regress and my realist interpretation are instantiations. The point of articulating such an archetype is that it enables me to make the case that the argument, once couched these in neutral terms, withstands objections that effectively rebut Korsgaard’s own argument. I allow that these objections meet their mark if the regress argument aspires to establish the truth of constructivism. But I argue that the argument succeeds if we recast it so that it purports to establish only that the exercise of rational capacities can sustain the final value of aims and that all agents are morally required not to violate the formula of humanity.

The central claims of my reconstruction of Korsgaard’s argument are as follows:

- (1) A complete explanation of the rationality of a person's performing an action must invoke the claim that some of her aims are valuable for their own sake.
- (2) If a complete explanation of the rationality of a person's performing an action must invoke the claim that some of her aims are valuable for their own sake, then a complete explanation of the rationality of her performing an action must invoke the claim that her rational capacities sustain the final value of some of these aims.
- (3) If a complete explanation of the rationality of a person's performing an action must invoke the claim that her rational capacities sustain the final value of some of her aims, then a complete explanation of the rationality of her performing an action entails that her possession of rational capacities entitles her to be treated only in ways that are consonant with recognition of the fact that these capacities can sustain value; let us say that anything entitled to this sort of treatment "merits respect".
- (4) If a complete explanation of the rationality of a person's performing an action entails that she merits respect by virtue of her possession of rational capacities, then a complete explanation of the rationality of her performing an action entails that any other individual with rational capacities also merits respect.
- (5) If a complete explanation of the rationality of a person's performing an action entails that any individual with rational capacities merits respect, then a complete explanation of the rationality of a person's performing an action entails that she is obligated not to violate the formula of humanity.
- (6) A complete explanation of the rationality of a person's performing an action entails that she is obligated not to violate the formula of humanity.

I believe that claims (1) through (6) can all be vindicated, and in Sections 5 and 6 below I defend these claims from within a realist value theory. But before doing so, I first need to explain why I reject Korsgaard's version of this argument, the regress argument.

Korsgaard has ambitions beyond those of my realist interpretation of the reconstructed argument, one of which is to present an argument that can rebut skeptical challenges. The regress argument is so-called

because it shares the same structure as putative cosmological proofs of the existence of God. These arguments posit a constraint on a complete explanation of causation—traditionally known as the principle of sufficient reason<sup>27</sup>—and seek to show that this constraint is satisfied only on the supposition that God exists. Korsgaard is clear that this is the structure of her argument: she uses the term “unconditioned condition”, an allusion to the canonical expression “unmoved mover”, and she explicitly analogizes her argument to pre-Kantian rationalist arguments for the existence of a first cause.<sup>28</sup>

As Korsgaard makes clear, and as is the case with any application of the principle of sufficient reason, the regress argument presupposes that a philosophical account of a fact is incomplete unless it rules out the possibility of that fact’s failing to obtain. More specifically, she claims that an account of the fact that a thing is valuable is incomplete if the account fails to preclude the obtaining of conditions that would defeat the claim that it is valuable.<sup>29</sup> By contrast, a complete explanation of the fact that a thing is valuable precludes all such defeating conditions, including any putative defeating conditions that might be advanced as skeptical hypotheses.

If a theory of causes is to rebut skeptical challenge by satisfying the principle of sufficient reason, it must posit a thing that is both the cause of all other things—including the cause of other things’ causal powers—and itself uncaused or self-caused; hence the characterization of this thing as an unmoved mover. Similarly, for a thing to play the role in Korsgaard’s regress argument that is structurally analogous to that of an unmoved mover in cosmological proofs of the existence of God, it must be a condition on the value of all other things and it must at the same time itself be unconditionally valuable.

#### **4. A Dilemma**

The trouble for this view, I argue in this section, is that there is no unconditioned condition of all value in this sense. The good will is not plausibly understood as the source of the value of human well-being; and while rationality is both a source of value and a source of moral requirements, it is not the only source of value and it is not the source of all value.<sup>30</sup> To see why, observe how Korsgaard’s exposition of her own view elides the distinction that I emphasize between the value of a good will and the value of human rational capacities. Once again I quote

somewhat extensively from Korsgaard's writings, in an effort to convey the spirit of her view. Here is Korsgaard in the opening chapter of *Creating the Kingdom of Ends*:

Kant's view is that only a good will has unconditional value of this kind. Since it is the objects of our own choices which we take to be good, and those objects do not have value in themselves, the source of value must be something that rests in us. It is not our needs and desires, for those are not always good. It must, therefore, be our humanity, our rational nature and our capacity for rational choice. *This is not different from saying it is a good will, for rational nature, in its perfect state, is a good will.*<sup>31</sup> [emphasis added]

In "Kant's Formula of Humanity":

To say that humanity is of unconditional value might seem, at first sight, somewhat different from the claim with which the Groundwork opens: that the good will is of unconditional value. *What enables Kant to make both claims without any problem is this: humanity is the power of rational choice, but only when the choice is fully rational is humanity fully realized.* Humanity ... is completed and perfected only in the realization of "personality", which is the good will. But the possession of humanity and the capacity for the good will, whether or not this capacity is realized, is enough to establish a claim on being treated as an unconditional end.<sup>32</sup> [emphasis added]

And, finally, in "Two Distinctions in Goodness":

On Kant's view there is only one thing that has what he calls unconditional value and what [G. E.] Moore calls intrinsic value, and that is the power of rational choice (*when the choices are made in a fully rational way, which is what characterizes the good will*).<sup>33</sup> [emphasis added]

In these passages, Korsgaard appears to deny the significance of distinguishing between the value of humanity and the value of the good will; but so far as I can tell, her writings do not settle definitively her view of the relationship between these two values.<sup>34</sup> The passage just quoted from "Two Distinctions in Goodness", with its qualification that rational choice has unconditional value only when it is exercised in a "fully rational way", suggests that humanity has unconditional value

only when it is fully rationally exercised as a good will. The passage from the opening chapter of *Creating the Kingdom of Ends* also suggests this understanding: saying humanity is unconditionally valuable could be “no different from saying” the good will is unconditionally valuable only if humanity is the same thing as a good will, and this could be the case only if humanity is not the mere capacity for rational choice but is instead the fully rational exercise of this capacity. The passage in the previous section from “Aristotle and Kant on the Source of Value” also suggests this interpretation. This passage states only that our aims have value “whenever they are chosen with full rational autonomy” and only that we act wrongly when we “act against their rational autonomy”, rather than claiming more generally that we act wrongly when we act against the (possibly imperfect) exercises of their rational capacities. By contrast, the passage from “Kant’s Formula of Humanity”, in stating that humanity and the capacity for a good will have unconditional value “whether or not this capacity is realized”, suggests that humanity has unconditional value even when not fully rationally exercised as a good will.

I do not know which understanding of the relationship between the value of humanity and the value of a good will is closer to Korsgaard’s own. But given the structure of her argument, it is crucial that she identify precisely the unconditioned condition of all value: is it humanity in general, or only humanity in its complete or perfect state? On this question, Korsgaard cannot elide the distinction between the value of a good will and the value of humanity.<sup>35</sup>

Whichever she chooses, however, her view will be unstable. Suppose she says humanity has unconditional value only when fully rationally deployed. This would lay to rest any worry of an explanatory gap between the value of humanity and the value of a good will: on this interpretation there is no such gap, since the extension of “humanity” in the statement of the formula of humanity is necessarily identical to the extension of “good will”. There is no need to assert a dubious bridge principle according to which potential Xs have all the value of fully realized Xs or imperfect Xs have all the value of perfect Xs.

The trouble with this interpretation is that it yields Kantian theories of value and obligation wildly at odds with our more specific judgments about value. Consider first the Kantian claim that the content of our moral obligations towards persons is captured by the formula of humanity. If we understand “humanity” as necessarily coextensive with

“good will”, then the formula of humanity fails to vindicate the claim that persons who lack a good will are worthy of respect. This is unappealing, to say the least, since one of the principal attractions of Kant’s theory is that it appears to have resources to explain the dignity of all rational beings.<sup>36</sup>

The results of this interpretation within the theory of value are, if anything, even more objectionable. The claim that the exercise of rational capacities can sustain the final value of aims is plausible, but is extremely controversial. The claim that the exercise of rational capacities can sustain the final value of aims, but only if this exercise manifests a good will, is not only controversial but implausible. A good will is not easy to achieve. Its presence is not guaranteed by the fact that one’s projects and relationships are permissible; nor is its presence guaranteed by the fact that all of the actions one takes in pursuing those projects and relationships are permissible. The presence of a good will entails, more strongly, that one would not pursue any aim if it were impermissible and that one would not pursue any aims impermissibly in any counterfactual circumstances where the goodness of one’s will is held constant.

These strong conditions on the presence of a good will explain why it is extremely difficult to know when a person exhibits a good will. And under the interpretation of “humanity” as necessarily coextensive with “good will”, this entails that it is extremely difficult to know when a project or relationship is valuable—whether one’s own or another’s—since only the projects and relationships of a good willer are valuable. But such a claim is false. The value of my own projects and relationships does not depend on any fact so rarefied as whether I have a good will; and it is no mistake to treat as valuable the permissible projects and relationships of those who lack a good will by virtue of their other, impermissible projects.<sup>37</sup>

I think these considerations are decisive against the view that humanity’s value is the value of humanity fully rationally deployed; but there is still the second horn of the dilemma to consider. On this interpretation, humanity is both unconditionally good and the source of the value of other goods, even when imperfectly exercised. It may have seemed obvious to some that this is the superior interpretation. As Korsgaard observes, however, this appears to contradict Kant’s claim that only a good will is good without qualification. More significantly, there is no way to motivate this claim analogous to the *Groundwork*

argument that a good will is good without qualification. It is obvious that humanity can be used for bad aims, for the “forces of evil”. Indeed some bad things, such as vicious actions, are necessarily constituted by the exercise of humanity. It is therefore implausible to think that humanity has the distinctive mode of value that Kant attributes to a good will. Rational capacities and their exercise are simply not, in general, good without qualification.

One lesson of this discussion is thus that the mode of value we should attribute to humanity in general is *not* the same as the mode of value—namely, being good without qualification—that Kant attributes to a good will. But if this is right, why attribute the mistake to Korsgaard? It would seem more charitable to read the passages connecting the value of humanity with the value of a good will as peripheral, and to recast the regress argument as proceeding independently from Korsgaard’s claims about the value of a good will.

The trouble with this suggestion is that Korsgaard is driven to elide the distinction between the value of humanity and the value of a good will by the structure and ambition of her regress argument. She needs her conclusion to be that humanity is unqualifiedly good because she is emulating cosmological arguments for the existence of God. Any qualification of the goodness of humanity—such as the obvious observation that rational capacities can be exercised viciously—would suffice to show that humanity is not the unconditioned condition of all value in a sense analogous to God’s role as unmoved mover in a cosmological argument. Since Korsgaard recognizes that impermissible aims are not valuable, even though their adoption is an exercise of humanity, she must invoke something other than the exercise of humanity itself to explain the difference between the value of permissible aims and the disvalue of impermissible aims. To maintain that humanity is the unconditioned condition of all value, in the face of this observation, would be akin to maintaining that something other than the unmoved mover’s activity must be invoked to explain the difference between cases in which the unmoved mover succeeds at causing events in the world and cases where it fails.

To concede such a point is to abandon the strategy of the cosmological argument, and Korsgaard is sensitive to the need to avoid making the analogous concession in her regress argument. I believe that this concern leads her to claim, to my mind implausibly, that saying humanity is unconditionally valuable is “not different” from saying that

a good will is unconditionally valuable, and that it leads her to claim that humanity is unconditionally valuable only “when the choices are made in a fully rational way”. Far from being sloppiness, the ambitions of Korsgaard’s argument push her strongly in the direction of the first horn of the dilemma, unappealing though it otherwise is. But our aims do not derive their value from a good will, and consequently we must reject the regress argument as Korsgaard understands it.<sup>38</sup>

One might object that the dilemma I have presented ignores a possibility that is intermediate between the claim that humanity is the unconditioned condition of all value and the claim that the good will is: namely, that *permissible* willing plays this role. I see two difficulties with this objection. One difficulty is that a permissible will is, as such, only accidentally permissible. There may be some counterfactual circumstance in which this will would falter; otherwise it would be a good will. This is a stark disanalogy with a cosmological argument, which would be vitiated under the hypothesis that God is only accidentally an unmoved mover.

A second difficulty with this objection is that the value of a permissible will is derivative, and consequently a permissible will cannot be the unconditioned condition of all value. There are two plausible understandings of why a permissible will is valuable: because it wills as a good will might, or because it is an exercise of humanity that is compatible with respect for all other exercises of humanity. Either way, a permissible will derives its value from something else—the value of a good will on the first account, and of humanity on the second—and so cannot itself be the source of all value.

## **5. Recovering Rational Capacities as Value-Sustaining**

My goals to this point have been largely negative: I have articulated a dilemma for Korsgaard’s regress argument in an effort to show that this argument should be rejected. I agree with Korsgaard, however, that claims (1) through (6) above can all be vindicated. I turn now to the positive goal of showing how that can be done. I begin with claim (1):<sup>39</sup>

(1) A complete explanation of the rationality of a person’s performing an action must invoke the claim that some of her aims are valuable for their own sake.

I assume that any behavior that is utterly unconnected to a person's aims is not a genuine exercise of rational agency. I assume further that unless the aims that make an action an exercise of rational agency have some value, the person's action cannot be rational. It need not be the case, of course, that all of a person's aims are valuable for their own sake; an aim of losing weight, for example, may be valuable simply because it is a means to better health. An explanation of the rationality of a person's aiming to lose weight still must invoke the claim that some of his aims are valuable for their own sake, however, since if there were no final value in health, or in something else health enables him to achieve, then he could not be rational in adopting and pursuing the aim of losing weight.

Note that claim (1) does not entail that every person judges that some of her aims are finally valuable; nor does it entail that every person endorses or otherwise has a pro-attitude towards her aims. It entails only that, to act rationally, a person must in fact have some aims that are finally valuable. Note also that claim (1) does not entail that every action is consciously chosen for the sake of or directed towards the achievement of some aim. It entails only that every action warrants the attribution of some aim, conscious or otherwise, to the person who performs it, and that either this aim or some more remote aim of the person must be finally valuable if she is to act rationally. I thus endorse literally a statement Benjamin Franklin made ironically: "It is so wonderful to be a rational animal, that there is a reason for everything that one does."<sup>40</sup> Franklin was poking fun at the human tendency to confabulate and to rationalize, to attribute to oneself false reasons and motivations; and no doubt we do invent reasons for our actions with regularity. But this truism does not undermine the claim that for a thing to be an action, it must be done for a reason. This truism also fails to undermine the claim that when we articulate the reason for which an action is performed, we attribute aims to the person who performed it, and it fails to undermine the claim that some of these aims must be finally valuable, if she is to act rationally.

The reconstructed argument continues with claim (2):

(2) If a complete explanation of the rationality of a person's performing an action must invoke the claim that some of her aims are valuable for their own sake, then a complete explanation of the rationality of her

performing an action must invoke the claim that her rational capacities sustain the final value of some of these aims.

This is likely the most controversial stage of the argument, and it is where my departures from Korsgaard's regress argument begin. I have already rejected her supposition that an explanation is complete only if it purports to rebut skeptical challenges by satisfying the principle of sufficient reason. I suggest instead that we abandon this putative constraint on satisfactory explanation and regard claim (2) as supported by an inference to the best explanation.<sup>41</sup> More specifically, I maintain that the hypothesis that the exercise of rational capacities can sustain the final value of choiceworthy aims is the best explanation of the conjunction of (i) the fact that it is rational when acting to treat actually adopted choiceworthy aims, but not non-adopted choiceworthy candidate aims, as valuable and (ii) the fact that it is rational to treat idiosyncratic features of a choiceworthy aim as valuable even when these features do not contribute to the aim's choiceworthiness. I turn now to the task of explaining how this inference proceeds and why it is warranted.

I start by attempting to show that we do not have, for each of our finally valuable aims, a complete account of why it is rational to treat that aim as finally valuable that is independent of its having been adopted as an aim. I start, that is, by arguing that we lack a wholly agent-independent explanation of the value of paradigmatically finally valuable aims, including interpersonal relationships (friendships and romantic relationships) and long-term projects (career and hobbies). Once I have argued for this claim, I seek to show how it provides resources that enable us to mimic the results of Korsgaard's regress argument from within a realist theory of value.

But first I must present the argument. Whenever I act, according to claim (1), I am rational only if some of my actual aims are finally valuable.<sup>42</sup> I thus treat the relevant aims as finally valuable when I act, since I must invoke the final value of these aims if I am to justify my action. I treat my vocation as a philosophy professor and avocation as a Scrabble enthusiast, for example, as though these aims generate reasons for action.<sup>43</sup> By contrast, I do not act as though there are reasons generated by the career I might have had, but do not have, as a politician, or by the hobby I might have had, but do not have, as a chess enthusiast.<sup>44</sup> Among the reasons I must posit if I am to act rationally are

reasons to adopt sub-aims of my actual aims, such as the sub-aims of writing essays and delivering lectures; but there need be no reasons for me to adopt sub-aims, such as the aims of raising funds and campaigning for office, of the aim I do not have of succeeding as a politician. These facts are familiar and undeniable: I treat my actual aims as valuable when I act, but I do not treat as valuable aims I have not adopted. Moreover, it is rational for me to do so, and the best explanation of this is that the adoption of aims can sustain their final value.

One might object to this claim by appealing to the putative fact that being a philosophy professor is more valuable for me than being a politician, or by appealing to the putative fact that being a Scrabble player is more valuable for me than being a chess player. Even if true, however, these latter claims play no role in explaining what I have reason to do, when I deliberate among candidate actions. To see why, suppose that I come to suspect that I should change my career or hobby. This entails that I have reason to deliberate about a possible change of career or hobby, but it does not entail that my actual aims cease to provide me with reasons for action. It may be rational for me to decide not to deliberate about these matters, or to postpone deliberation. If it is rational for me to postpone or cancel deliberation, then it is rational for me to continue to treat my actually adopted career and hobby as generating reasons for action—assuming these aims are sufficiently valuable for me to pursue, even if they are by hypothesis not the *most* valuable—just as I did before I suspected a change might be in order.

Or suppose I come to realize that I have greater aptitude as a politician than as a philosophy professor, or I come to realize that I live in a society that values politics more highly than philosophy.<sup>45</sup> Even then my actually adopted aim determines what I have reason to do, in the context of deliberation among actions, provided that it is sufficiently valuable. To explain this we need to look beyond the value my aims have independent of my adopting them—their “choiceworthiness”<sup>46</sup>—and advert to a fact about how I am exercising my rational agency. A complete explanation of why it is rational for me to treat succeeding as a philosophy professor as valuable, but not succeeding as a politician, must include the fact that I aim to succeed as a philosophy professor and the fact that I do not aim to succeed as a politician.

These considerations show that the reasons I have for performing this or that action are sensitive to the projects and relationships I have

actually adopted. They do not yet show, however, that my adoption and pursuit of aims—the exercise, that is, of my rational capacities—sustains their final value. One might think against this that there are general features of human agency that completely explain the rationality of treating only actually adopted aims as valuable when deciding how to act. I am a finite being, and so I cannot entertain every consideration that bears on what I should do; nor can I pursue all the aims that are worthwhile for me to pursue. I must organize the exercise of my agency by adopting aims, and I must temporally and hierarchically order my pursuit of these aims.<sup>47</sup> One might think that candidate aims have whatever value they have, and that it is only a consequence of my finitude that when I act I am rational to treat some of these candidate aims as valuable (the choiceworthy ones I have actually adopted) but not others (those I have not adopted or are not choiceworthy).

But this is not right. To see why, and so to complete the argument for claim (2), we must focus not on features of our aims that make them worthy of choice, but instead on their idiosyncratic features that do not contribute to their choiceworthiness. My pursuit of Scrabble as a hobby, for example, is choiceworthy because Scrabble expertise consists in mental skills of wide application, such as facility recalling information from memory, large vocabulary, and quickness in the performance of arithmetic. But Scrabble expertise also consists in part in arcane skills that are far less useful, such as knowing all the words of English that contain a “Q” but not a “U” and having familiarity with the details of the *Official Scrabble Player’s Dictionary*. If the value for me of pursuing Scrabble is exhausted by whatever value this aim has independent of my adoption of it, then all that I should value about my hobby are the features that make it choiceworthy, such as the fact that it enables me to develop and exercise my cognitive capacities. In particular, I should not treat the idiosyncrasies of Scrabble as themselves helping to constitute the hobby’s value, and this should in turn give me reason to regret that my pursuit of Scrabble leads me to develop and exercise these cognitive capacities in a peculiarly Scrabble-like way. Under this hypothesis, I should regard peculiarities of Scrabble as unfortunate, as impurities or inefficiencies of the hobby as a way to develop and exercise cognitive capacities.<sup>48</sup>

But this is false to the facts of what it is like to pursue a hobby like Scrabble. For the Scrabble enthusiast, the idiosyncrasies of the game are, in the context of deliberations about what to do, on a par with the

features that make it choiceworthy; *and this does not appear, on reflection, to be a mistake*. If the entire activity of Scrabble-playing consisted in deploying arcane knowledge, Scrabble would be a relatively unchoiceworthy aim. But since Scrabble expertise is, we may stipulate, a choiceworthy aim for the enthusiast, the peculiar Scrabble-reasons have no second-class status; there is nothing to be regretted in valuing success in the idiosyncratic features of our choiceworthy aims.

The best explanation of this is that Scrabble and other choiceworthy aims have a mode of value that is sustained by their adoption. Scrabble is choiceworthy—it has value for me independent of my adoption of it, and there is some reason for me to adopt it as an aim—by virtue of the fact that it constitutes the exercise of cognitive capacities and is a means to the improvement of these capacities. But once I adopt Scrabble as a hobby, it is also rational for me to treat success in Scrabble as valuable for its own sake. That is why success in its idiosyncratic features is valuable for me. It is thus rational to treat the pursuit of Scrabble by those who have adopted Scrabble-playing as valuable for its own sake, but it is not rational to treat the pursuit of Scrabble by those who have not adopted this aim as valuable in that way. Since the only relevant point of difference these two groups of individuals is that the former have adopted the aim and the latter have not, the best explanation of these facts is that adopting the aim sustains a part of its value—its final value—that it would not otherwise have.

Thus far I have used a hobby to illustrate this inference. But if the preceding considerations motivate the claim that the exercise of rational capacities can sustain the final value of hobbies, they also motivate the claim that the exercise of rational capacities can sustain the final value of other sorts of projects. Arcane insect knowledge is finally valuable for an entomologist, but not for others.<sup>49</sup> The claim also applies to interpersonal relationships. It is rational to value actual friendships and romantic relationships for their own sake; it is a gross error to value a friendship only as a means to happiness or as constituting a way of having *a* friend. My suggestion is that these facts are explained by the hypothesis that adopting the relationship as an aim generates its final value.<sup>50</sup>

I thus contend that the best explanation of why it is rational for us to treat our actually adopted aims for their own sake, while it is not rational for us to value other choiceworthy candidate aims for their own sake, is that the exercise of rational capacities sustains the final value of

choiceworthy aims. This argument for claim (2) does not share the motivations behind Korsgaard's constructivism, and so it may be advanced by a value realist. Furthermore, its conclusion—its interpretation, that is, of claim (2)—is much weaker than Korsgaard's, since the argument does not purport to establish that *all* value is sustained by the exercise of rational agency. Indeed, the argument presupposes that some value is *not* sustained in this way, since it deploys a notion of choiceworthiness that applies to candidate aims independent of whether they are adopted.<sup>51</sup> This notion of choiceworthiness figures crucially in the best explanation of the value of our projects and relationships, and is not dispensable in favor of a thoroughgoing constructivist account of the value of these aims.

This exposes another respect in which my argument is weaker than Korsgaard's: I do not assert that rational capacities can generate value from nothing, since I claim that only *choiceworthy* aims can have final value when they are adopted. Independent of my adoption of it as an aim, Scrabble is non-finally valuable as a way to constitute both a hobby and the exercise of cognitive capacities. Independent of my adoption of it as a career, being a philosophy professor is non-finally valuable as a way to constitute both a career and the exercise of cognitive capacities. Independent of my actually developing relationships with my friends, candidate friendships are non-finally valuable as a way to constitute a friendship and as a way to exercise emotional capacities. When I adopt these aims, however, I claim that they acquire a mode of value—their final value—that they did not previously have.

It is important to observe that the above argument does not seek to show merely that we are, by virtue of our possession of rational capacities, able to generate things (our aims) of final value that did not exist before. That might establish that these capacities are instrumentally valuable, but it would not show that they have a mode of value that supports the formula of humanity. To see why we should endorse the stronger claim, we should focus not only on how the exercise of rational capacities helps generate the final value of a choiceworthy aim when it is adopted, but also on how the exercise of rational capacities sustains the final value of our choiceworthy aims over time. We can change what is finally valuable by changing our aims; and when we abandon an aim, it typically ceases to be finally valuable. Our aims thus depend for their final value on the continued activity of our agency.<sup>52</sup>

## 6. Recovering the Formula of Humanity

Thus far I have argued that rational capacities can sustain the final value of choiceworthy aims. If we accept this claim, the reconstructed argument continues with its assertion of claim (3), then we should regard those with rational capacities as worthy of respect:

(3) If a complete explanation of the rationality of a person's performing an action must invoke the claim that her rational capacities sustain the final value of some of her aims, then a complete explanation of the rationality of her performing an action entails that her possession of rational capacities entitles her to be treated only in ways that are consonant with recognition of the fact that these capacities can sustain value; let us say that anything entitled to this sort of treatment "merits respect".

Claim (3) is used to establish that the capacity to sustain the final value of aims is itself a mode of value.<sup>53</sup> If this right, then we should recognize this mode of value and respond appropriately to it. It is a further and more controversial claim to assert that the rational response to this mode of value is Kantian respect, which consists in such things as stable dispositions to refrain from coercing, deceiving, or destroying the object of respect and in a stable disposition to treat the happiness of the object of respect as valuable. It would take us too far afield to investigate these Kantian claims about the substance of our moral requirements, although a complete defense of the formula of humanity would of course have to establish them. The claim here is only that there is some substantive requirement, expressed by the formula of humanity, to respect anything with the capacity to sustain final value.<sup>54</sup>

The reconstructed argument proceeds with the observation that there is nothing about the exercise of one's own rational capacities that should make it differ in value from others'. It infers from this that we must treat others' humanity as meriting of the same sort of respect as our own:

(4) If a complete explanation of the rationality of a person's performing an action entails that she merits respect by virtue of her possession of rational capacities, then a complete explanation of the rationality of her performing an action entails that any other individual with rational capacities also merits respect.

This stage of my realist interpretation of the reconstructed argument differs crucially from Korsgaard's regress. Much criticism of the regress focuses on Korsgaard's attempt to move from the claim that one's own humanity merits respect to the claim that the humanity of others also merits respect; she has been charged with begging the question against an egoist willing to entertain skeptical hypotheses that would vindicate the view that only *his* rational capacities are a source of value. This makes sense, in view of the ambitions of the regress argument. As we have seen, Korsgaard purports to establish not only that this inference works, but more strongly that it is an application of the principle of sufficient reason which can rebut such skeptical challenges. That is a tall order, and Korsgaard's commentators have appropriately questioned whether she pulls it off.<sup>55</sup>

This stage of the argument can also be recast, however, as an inference to the best explanation; and once we do this, I contend, my realist interpretation of the reconstructed argument avoids these objections to the regress argument. In fact, once it is recast as an inference to the best explanation, this stage of the argument is wholly innocuous. It would be patently absurd for an egoist to claim that his rational capacities can sustain the final value of his aims but to deny that your rational capacities can sustain the final value of your aims, if he supports the former claim with an inference to the best explanation of the final value of his actual aims.<sup>56</sup> This would be a mistake in explanatory reasoning akin to postulating without motivation that while all observed electrons have a negative charge that explains their attraction to protons, there are other electrons that lack charge and so are not attracted to protons. What, after all, is the more plausible explanatory hypothesis: that rational capacities in general are able to sustain the final value of actual choiceworthy aims, and that as a consequence the egoist's rational capacities are able to do this; or that the egoist's rational capacities alone have can sustain value in this way, and that consequently his aims are finally valuable even though others' are not? An egoist who resists the first hypothesis is reduced to the posture of a solipsistic skeptic who stubbornly resists the hypothesis that the behavior of other persons is explained by the fact that they have minds.<sup>57</sup>

If we could treat as a relatively fixed point of inquiry that the aims of others *lack* value, then this inference would be far from obvious. It would then be more reasonable for the egoist to posit that there is

something special about his rational capacities. His position would be like that of a physicist who observes electrons that, for unexplained reasons, fail to be attracted to protons; given such a data set, it might well be reasonable to attribute negative charge to some electrons but not to others. But it is not reasonable to begin from an unmotivated skepticism about the value of others' aims. Even if the egoist supposes only that he is agnostic about the value of others' aims, the inference to the fact that his rational capacities can sustain the final value of his aims should lead him to infer that other people's rational capacities can sustain the final value of their aims.

Notice also that it is the very same explanatory inference that vindicates both the claim that one's own rational capacities can sustain the final value of choiceworthy aims and the claim that others' rational capacities can do this. On my reconstructed argument, then, it is misleading to think of the argument for the formula of humanity as proceeding in two stages, the first of which seeks to establish that I merit respect and the second of which extends this status to other persons. The same explanatory hypothesis that is able to vindicate the rationality of attributing final value to my aims—the hypothesis that rational capacities can sustain the final value of choiceworthy aims—vindicates the claim that all other persons merit respect.

From here it is a small step to the remaining claims of the argument:

5) If a complete explanation of the rationality of a person's performing an action entails that any individual with rational capacities merits respect, then a complete explanation of the rationality of a person's performing an action entails that she is obligated not to violate the formula of humanity.

This claim is a gloss of what violations of the formula of humanity consist in. As I indicated above, I will not here attempt to argue that the details of Kant's understanding of the formula of humanity are supported by my reconstructed argument; but I do believe that something like this is the case. And from claims (1) through (5) it follows, as a matter of deductive logic, that:

(6) A complete explanation of the rationality of a person's performing an action entails that she is obligated not to violate the formula of humanity.

We are required to heed the formula of humanity because any failure to do so treats humanity in one person (oneself) as meriting respect and simultaneously treats humanity in another person (possibly but not typically oneself) as lacking this mode of value.

Freed from the ambitions of Korsgaard's regress—vindicating her thoroughgoing constructivism and rebutting skeptical challenges—we can mimic her justification of the formula of humanity. Each person treats some of their aims as finally valuable, since some of these aims must be finally valuable for her actions to be rational; but the best explanation of why it is rational to treat one's aims as finally valuable invokes the claim that rational capacities can sustain the final value of these aims; each person therefore must treat rational capacities as though they merit respect; and this is true whether the rational capacities in question are one's own or those of another. Mimicking Korsgaard in this way enables a value realist to vindicate the constructivist's route to a moral requirement not to violate the formula of humanity without endorsing constructivism's claims about the nature of value itself.

## **7. Two Kinds of Unconditional Value**

In this final section I briefly sketch the account of morality that emerges from my realist interpretation of the reconstructed argument. The argument, if successful, shows that anyone who violates the formula of humanity makes an exception of himself by treating his own rational capacities as though they merit respect while treating some person's rational capacities as though they lack the same value.<sup>58</sup> This is an important feature of common sense understandings of wrongful action, which is shared by other Kantian moral theories, and which helps explain how failing to comply with the formula of humanity differs from other sorts of rational failure.

My realist argument is able to capture this appealing feature of Kantian moral theory because, like Korsgaard's constructivism, it explains how the formula of humanity can regulate the pursuit of values even though the requirement not to violate the formula of humanity emerges from within the theory of value. In my view, this is the greatest appeal of constructivism: it makes moral requirements less mysterious by integrating claims about them into a broader theory of value.<sup>59</sup> The chief aim of this essay is to show how to do this without endorsing a

thoroughgoing constructivism about value and thereby incurring that view's attendant difficulties.

I say that this makes moral requirements less mysterious because it helps explain how and why they regulate action. More specifically, it helps explain why morality is pervasive and overriding.<sup>60</sup> To say that morality is pervasive is to say that there are no morality-free zones within which moral norms do not apply. This thesis is supported by my realist interpretation of the reconstructed argument because it purports to show that a requirement to respect humanity is involved in every action a person performs. To say that morality is overriding is to say that a verdict that an action is morally required cannot be defeated by values that might be realized by refraining from performing the action. This thesis is supported by my realist argument because it purports to show that some of the value we attribute to our aims is sustained by the exercise of our rational capacities, and so depends on the value of those capacities. This is only the beginning, of course, of a defense of the overridingness of morality. There is no straightforward entailment of overridingness by the claim that rational capacities can sustain the final value of aims. But this claim provides support for the claim of overridingness, since it shows that the value that underlies the existence and force of our moral obligations towards other people is internal to the rationality of pursuing aims for their own sake.

This understanding of morality involves attributing a special value to rational capacities. On this account, rationality is a good in itself, a value that pervades all exercises of practical reason and overrides the mode of value we attribute to our finally valuable aims.<sup>61</sup> This sounds like the usual Kantian attribution of unconditional value to humanity; but did we not already observe serious difficulties with such an attribution?

What was previously argued is that we should not claim that rational capacities have the same mode of value that Kant attributes to a good will, and that we should not claim that humanity's value derives from the value of a good will. Rejecting these claims does not entail, however, that there is no significant sense in which rational capacities are unconditionally good. It entails only that if they are unconditionally good, then there are at least two ways in which a thing can be unconditionally good. And this, I think, is exactly what we ought to say.

One sense of "unconditionally good" is that of "good without qualification" or "good without limitation": unqualified goods are those, roughly speaking, that cannot be used for the forces of evil. This is the

most rarefied form of goodness, and only a good will is plausibly held to be good in this way.<sup>62</sup> Another sense of “unconditionally good” is that which Kantians express with the terms “good in itself” or “end in itself”: these are goods whose value does not depend on the value of anything else, and whose goodness is present in all conditions, even if sometimes—or always—in a qualified way. On the use that I recommend, this sort of goodness counts as “unconditional”, for if a thing is good in itself—if its value does not depend on the value of anything else—then it is good in all conditions.<sup>63</sup>

I believe that this sort of value—being an end-in-oneself—tracks the category of *moral standing*; it captures all and only those things that one must take account of in order to be a good person. And in closing, although I do not have space to explore the proposal here, I would observe an ancillary benefit of using my reconstructed argument to support the formula of humanity. This argument disconnects the value of rational capacities from the value of a good will, and so opens the door for the straightforward attribution of moral standing to beings that lack the capacity for a good will, such as infants and intelligent animals. A new argument would need to be advanced to establish this, one that motivates, for all persons, a requirement to attribute value to the capacity of sentience as well as to rational capacities. That argument must wait for another occasion, but the conceptual space to pursue it is opened by the strategy for mimicking Korsgaard I have presented here.

---

<sup>1</sup> I would like to thank Ben Chan, Brad Cokelet, Louis DeRosset, Kyla Ebels Duggan, Sam Fleischacker, Barbara Herman, Richard Kraut, Laura Papish, Julie Tannenbaum, Melissa Yates, and Rachel Zuckert for their feedback on earlier drafts of this essay. I would also like to thank the audience members at Purdue University for the 2008 Midwest Study Group of the North American Kant Society—including especially Brian Chance, Patrick Kain, Jeremy Schwartz, Daniel Sutherland, Kristi Sweet, and Allen Wood—for their excellent questions.

<sup>2</sup> Korsgaard (1996a, 1996b).

<sup>3</sup> Michael Smith (1999), 385.

<sup>4</sup> Allan Gibbard (1999), 140.

<sup>5</sup> Allen Wood (1998b), 607.

<sup>6</sup> As I use the term, a constructivism is “thoroughgoing” just in case it is incompatible with realism. This distinguishes Korsgaard’s constructivism from the view of Allen Wood, which is similar but contends that there is one realist value, namely, humanity. See Wood (1999). This also distinguishes Korsgaard’s view from that espoused in this essay, which some might label constructivist even though it allows a wide range of realist values.

<sup>7</sup> It is now common to read Kant as a constructivist, following the influential interpretations of John Rawls (1980), Korsgaard (1986a), and J. B. Schneewind (1991). This reading is not, however, uncontroversial; for an important dissent see Karl Ameriks (2003), especially Chapter 11.

<sup>8</sup> For such a systematic defense of non-naturalist realism, see Russ Shafer-Landau (2003).

<sup>9</sup> For discussions of Korsgaard’s regress argument see Berys Gaut (1997), William FitzPatrick (2005), and Michael Ridge (2005).

<sup>10</sup> I am largely sympathetic to the criticisms – some offered in a friendly spirit, others not – advanced by Gaut (1997), FitzPatrick (2005), and Ridge (2005). I do not discuss their criticisms in detail here.

<sup>11</sup> Kant (1785), 7, or G393.

<sup>12</sup> Kant (1785), 7-8, or G393-394.

<sup>13</sup> Kant (1785), 38, or G429.

<sup>14</sup> There is a large literature concerning how to understand Kant’s notion of humanity, which he expresses with his technical use of the German word “humanität”. I do not engage that discussion here, since the details of Kant’s own account of humanity should not determine the success or failure of the arguments I present here. A full treatment of this issue would need to explore the relationships between humanity and various human capacities, including sentience, the capacity to obey the moral law, the capacity to value things, and the capacity to adopt and pursue aims. This fuller treatment would need to engage not only Kant’s discussions of humanity in the *Groundwork* but also his uses in other texts, including especially his *Religion Within the Limits of Reason Alone*. See Kant (1793), especially Book One.

---

<sup>15</sup> For more on this point see Wood (1998a).

<sup>16</sup> See Kant (1785), 19-21, or G406-408; and Kant (1793), 27-34. In this passage from the *Religion*, Kant appears more strongly to deny that any human has ever manifested a good will, and this stronger view is also at least suggested by the passage from *Groundwork II*.

<sup>17</sup> Samuel Kerstein suggests, to my mind implausibly, that we attribute this view to Kant. For his ingenious defense of this position, see Kerstein (2002, 2006).

<sup>18</sup> See Kant (1793), 27-33.

<sup>19</sup> See Kerstein (2006) for an instance of this strategy, albeit an idiosyncratic one.

<sup>20</sup> Here I echo Wood (1999), 120-121.

<sup>21</sup> See Korsgaard (1983, 1986a, 1986b, 1996a, 1996c, 2002, 2003).

<sup>22</sup> Korsgaard (1996c), 16.

<sup>23</sup> Korsgaard (1986a), 122-123.

<sup>24</sup> Korsgaard (1986b), 240-241.

<sup>25</sup> Korsgaard (1983), 258-259.

<sup>26</sup> Once again, for more complete discussions of Korsgaard's own argument, see Gaut (1997), FitzPatrick (2005), and Ridge (2005).

<sup>27</sup> Like canonical users of the principle of sufficient reason, Korsgaard intends her regress argument to be purely apriori justified. This is among the ambitions of her argument that I abandon here.

<sup>28</sup> Korsgaard (1983), 259.

<sup>29</sup> It may appear surprising that we should attribute an argument of this structure to Kant. As Korsgaard notes, in the *Critique of Pure Reason* Kant famously criticizes his rationalist predecessors' deployment of the principle of sufficient reason to try to establish the existence of a first cause. But it is of course open to Kant to claim that the principle of sufficient reason can be deployed in the theory of value even though not in the theory of causes, on the ground that we have the ability to cognize an unconditioned condition of value – namely, a good will – but lack the ability to cognize an unmoved mover. Korsgaard endorses this idea: "Theoretical reason, in its quest for the unconditioned, produces antinomies; in the end, the kind of unconditional explanation that would fully satisfy reason is unavailable. Practical reason in its quest for justification is subject to no such limitation." See Korsgaard (1986a), 119.

<sup>30</sup> As I explain below in Section 5, the sense in which I claim that rational capacities can help sustain the final value of aims is attenuated by comparison with the sense in which Korsgaard claims this. Also, as I suggest in Section 7, I do not believe, as Kant did, that the content of our moral obligations is exhausted by the formula of humanity; I contend against this that sentience is also a locus of moral standing.

<sup>31</sup> Korsgaard (1996c), 17.

<sup>32</sup> Korsgaard (1986a), 123-124.

---

<sup>33</sup> Korsgaard (1983), 262.

<sup>34</sup> For further discussion of Korsgaard's views on this issue, see Gaut (1997) and Kerstein (2006).

<sup>35</sup> It could be that she relies on an implicit bridge principle of the sort considered above, which appeals to the potential for humanity to manifest a good will. But since she here emulates cosmological arguments, this strategy would be an instance of the first horn of the dilemma, as it derives the value of humanity from the unconditional value of a good will.

<sup>36</sup> In a highly idiosyncratic interpretation, Richard Dean attributes this view to Kant; see Dean (2006). For a criticism of Dean's interpretation, see Patrick Frierson (2007). I sympathize with Frierson's criticisms, and whatever the merits of attributing the view to Kant, the considerations in this paragraph and the next are in my view decisive against this view on the merits.

<sup>37</sup> Kerstein suggests that the presence of a good will may be compatible with the presence of impermissible aims; see Kerstein (2006). I oppose this understanding of what it is to have a good will, but I do not discuss the issue here.

<sup>38</sup> Although I do not investigate the issue here, Korsgaard's argument in her Locke lectures for the view known as "constitutivism" will not save the regress argument; see Korsgaard (2002). Very briefly: even if we suppose (as does not seem likely) that all rational beings respond to the categorical imperative, it is not the case that they all do so in a way that manifests a good will. We can therefore distinguish between those exercises of rationality that manifest a good will and those do not, and generate the dilemma for the regress argument by asking which of these categories is held to be the unconditioned condition of all value. If it is claimed that all rational beings *do* manifest a good will, this expands the extension of "good will" in a way that undermines the *Groundwork* I arguments that a good will is good without qualification.

<sup>39</sup> Compare this claim to Korsgaard in *The Sources of Normativity*: "Since you are human you *must* take something to be normative, that is, some conception of practical identity must be normative for you. If you had no normative conception of your identity, you could have no reasons for action, and because your consciousness is reflective, you could then not act at all." See Korsgaard (1996a), 123. I disagree with Korsgaard's view that reflective endorsement is the mechanism by which agents can sustain the final value of their aims, but I do not investigate that issue here.

<sup>40</sup> United States Supreme Court Justice David Souter attributed this quip to Franklin during oral arguments for the case *United States v. Lopez*, which concerned the authority of the federal government to regulate firearms in schools. For a transcript see [http://www.oyez.org/cases/1990-1999/1994\\_93\\_1260/argument/](http://www.oyez.org/cases/1990-1999/1994_93_1260/argument/).

---

<sup>41</sup> Ridge (2005) labels Korsgaard's own argument as an inference to the best explanation, but this is either mistaken or misleading. In some sense almost any argument involves an inference to the best explanation, and deployments of the principle of sufficient reason in particular appeal to standards of explanation to justify their moves. But it is misleading to label a deployment of the principle of sufficient reason as an inference to the best explanation, since it is common practice to reserve the latter label for an argumentative form that resists the high standards of putatively skeptic-rebutting rationalism. Korsgaard's argument is not an inference to the best explanation in this sense.

<sup>42</sup> I am agnostic here regarding the order of explanation between reasons and values. My own view is that values are more basic, but that view is not needed to motivate the arguments in the text.

<sup>43</sup> I reiterate here that an agent's treating an end as valuable does not, in general, entail that the he endorses the end, would endorse the end on reflection, or that he is disposed to have any conscious pro-attitude towards it at all. A person can question whether she is right to treat her ends as valuable and a person can believe that it would be better, all things considered, if she abandoned or revised some of her ends; indeed, a person can judge some of her ends neutral or even negative in value. But these caveats do not undermine the claim that insofar as an agent has an end she *treats* it as valuable. An alcoholic treats drinking as valuable when she procures and imbibes, since she shapes her actions as though drinking were positively valuable, even if she judges on the whole that it is negatively valuable.

<sup>44</sup> To be more precise: the only sorts of reasons they provide are reasons to stop deliberating among actions and start deliberating among aims.

<sup>45</sup> Note that it might, under these circumstances, be rational for me to abandon my aim of succeeding at philosophy and adopt the aim of succeeding at politics; but that is a different claim.

<sup>46</sup> I adapt this notion of choiceworthiness from T. M. Scanlon; see Scanlon (1998), 131.

<sup>47</sup> For extensive developments of this claim see Rawls (1971), especially Part III, and see also Michael Bratman (2007).

<sup>48</sup> I will be rational to regard Scrabble's idiosyncrasies as impurities insofar as I regard Scrabble itself as merely constitutively valuable, that is, as valuable merely because playing Scrabble constitutes the exercise of my cognitive capacities; and I will be rational to regard these idiosyncrasies as inefficiencies insofar as I regard Scrabble itself as merely instrumentally valuable, that is, merely as a means to the improvement of my cognitive capacities.

<sup>49</sup> It is perhaps plausible that all knowledge is finally valuable, and that as a consequence all arcane insect knowledge is finally valuable. This does not threaten the argument in the text, however, since a salient difference between the

---

two cases remains: an entomologist has a much stronger reason to value arcane insect knowledge than another person does.

<sup>50</sup> Note that this is *not* an endorsement of the view, advanced by David Schmidtz and Harry Frankfurt, that appeal to instrumental reasons for treating an aim as finally valuable can fully rationalize treating the aim this way. See Schmidtz (1994) and Frankfurt (2004), especially Section 10 of Chapter Two. I cannot argue fully against this view here, but I would briefly mention a few reasons for dissent: valuing an aim finally for instrumental reasons is unstable, since reflecting on how one came to adopt the aim undermines the conviction that it is finally valuable; valuing an aim finally for instrumental reasons makes the generation of final value too mysterious and too volitional, since it opens widely the conditions under which persons can sustain the final value of aims; and this model assimilates paradigmatic cases of rationality, such as treating choiceworthy projects and relationships as finally valuable, to cases of borderline irrationality, such as valuing a higher power for its own sake in order to stay sober or valuing a diet for its own sake in order to lose weight.

<sup>51</sup> The realist interpretation of the reconstructed argument is thus able to avoid Stephen Darwall's principal objection to the regress argument; see Darwall (2006), 230-231.

<sup>52</sup> This is not a wholly voluntary ability. Some aims may be so deeply integrated into our life-plans that it is extremely difficult rationally to cease treating them as finally valuable, and as I have noted aims must be at least minimally choiceworthy if we are to sustain their final value. But these caveats do not undermine the argument in the text.

<sup>53</sup> I endorse Allen Wood's claim that "... rational nature is not being viewed as the source of *good things* (i.e., of their *existence*), but instead as the source of the fact of their *goodness*", if we qualify this claim so that the goodness in question is the final value of choiceworthy aims. As Wood maintains (and Korsgaard as well), being a source of value in this sense is itself a mode of value, and this mode of value is not merely instrumental. See Wood (1999), 130; the emphases are Wood's.

<sup>54</sup> One could object that even if the realist interpretation of the reconstructed argument is able to establish the irrationality of violating the formula of humanity, it fails to establish an *obligation* not to violate the formula, and fails to establish an entitlement against others' violations of the formula. This objection might maintain that the argument runs afoul of Stephen Darwall's dictum "second-personal authority out, second-personal authority in" or that it illicitly infers claims of what Michael Thompson calls "dikaiological" form from claims of what he calls "monadic" form. See Darwall (2006), 57-60, and Thompson (2004). It would take us too far afield to investigate these issues here. This essay attempts to show that there is a categorical rational requirement not

---

to violate the formula of humanity; a full defense of the claim that this requirement is best understood as an obligation must be taken up elsewhere.

<sup>55</sup> See, in particular, Gibbard (1999) and Ridge (2005).

<sup>56</sup> This answers Stephen Darwall's objection to the regress argument at Darwall (2006), 231-233.

<sup>57</sup> Although my route to the conclusion is very different from his, I endorse – word for word – the following remark by David Velleman: “Moral philosophers have long struggled with the problem of pinpointing the immoralist’s inconsistency, since he certainly isn’t guilty of any obvious self-contradiction or deductive fallacy. My proposal is that the immoralist’s inconsistency can be diagnosed by the standards of inductive logic, as formulated in our canons of theory-choice. The immoral agent, in his treatment of persons, makes himself a special case, an exception; and although he doesn’t thereby transgress any law of deductive inference, he does transgress a principle of theory-choice – namely, the principle that for the sake of generality and simplicity, a theory should avoid having to provide for special cases and exceptions.” See Velleman (2007), 306. For a related view, see also FitzPatrick (2005).

<sup>58</sup> This way of understanding morality helps explain the connection between Kant’s formula of humanity and his formula of universal law. For an allied view, see the quote from David Velleman in note 57.

<sup>59</sup> Korsgaard would agree that this is a virtue of constructivism, but would claim that constructivism also helps to resolve long-standing questions about the metaphysics and epistemology of value. I do not engage with those issues directly here.

<sup>60</sup> For an excellent exposition of these notions, see Samuel Scheffler (1992), especially Chapter Two.

<sup>61</sup> The claim that there are different modes of value, and that these different modes of value can call for distinct appropriate responses, is an important theme of the work of Elizabeth Anderson; see in particular Anderson (1993). This claim is developed richly in the context of Kantian moral theory by Barbara Herman (1981, 1983, 1984, 1985, 1993, 2001); see also Marcia Baron (1995), and Wood (1999, 2008). I concur with these theorists that we should reject the premise, implicit in much literature in both philosophy and the social sciences, that something like G. E. Moore’s conception of value is correct. On this conception of value, value appends in the first instance to states of affairs; anything else that is of value is so by virtue of its capacity or tendency to produce states of affairs that are independently specifiable as valuable. This understanding of value encourages the idea that the appropriate response to value is production; on this understanding, a rational person is one who is able to produce the most valuable states of affairs available in the circumstances. But as these theorists have argued, production is not the only fitting response to value. An important alternative response to value for our purposes is respect;

---

part of my proposal is that we cash out the Kantian notion of meriting respect, in the style of Korsgaard, in terms of having a value-sustaining status. To respect a person is to treat that person as having this status by virtue of his capacity to sustain the final value of his choiceworthy aims.

<sup>62</sup> I take no position here on whether a good will is, in fact, good in this way.

<sup>63</sup> For a related proposal, see Rae Langton (2007).

---

 Works Cited

- Ameriks, Karl (2003). *Interpreting Kant's Critiques*. Oxford, 2003.
- Anderson, Elizabeth (1993). *Value in Ethics and Economics*. Harvard, 1993.
- Baron, Marcia W. (1995). *Kantian Ethics Almost Without Apology*. Cornell, 1995.
- Bratman, Michael (2007). *Structures of Agency*. Oxford, 2007.
- Darwall, Stephen (2006). *The Second-Person Standpoint*. Harvard, 2006.
- Dean, Richard (2006). *The Value of Humanity in Kant's Moral Theory*. Oxford, 2006.
- FitzPatrick, William J. (2005). "The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism, and the Nature of Normativity". *Ethics* 115:4, 2005.
- Frankfurt, Harry (2004). *The Reasons of Love*. Princeton, 2004.
- Frierson, Patrick (2007). Review of *The Value of Humanity in Kant's Moral Theory*. Notre Dame Philosophical Reviews, 2007.
- Gaut, Berys (1997). "The Structure of Practical Reason". In *Ethics and Practical Reason*. Eds. Garrett Cullity and Berys Gaut. Oxford, 1997.
- Gibbard, Allan (1999). "Morality as Consistency in Living: Korsgaard's Kantian Lectures". *Ethics* 110:1, 1999.
- Guyer, Paul (2000). *Kant on Freedom, Law, and Happiness*. Cambridge, 2000.
- Herman, Barbara (1981). "On the Value of Acting from the Motive of Duty". *Philosophical Review* 90, 1981.
- \_\_\_\_\_ (1983). "Integrity and Impartiality". In *The Practice of Moral Judgment*. Harvard, 1993.
- \_\_\_\_\_ (1984). "Mutual Aid and Respect for Persons". *Ethics* 94, 1984.
- \_\_\_\_\_ (1985). "The Practice of Moral Judgment". *Journal of Philosophy* 82, 1985.
- \_\_\_\_\_ (1993). "Leaving Deontology Behind". In *The Practice of Moral Judgment*. Harvard, 1993.

- 
- \_\_\_\_\_ (2001). "The Scope of Moral Requirement". *Philosophy and Public Affairs* 30:3, 2001.
- Kant, Immanuel (1785). *Groundwork of the Metaphysics of Morals*. Trans. and Ed. Mary Gregor. Cambridge, 1997.
- \_\_\_\_\_ (1793). *Religion Within the Limits of Reason Alone*. Trans. and Eds. Theodore Greene and Hoyt Hudson. Harper, 1934.
- Kerstein, Samuel (2002). *Kant's Search for the Supreme Principle of Morality*. Cambridge, 2002.
- \_\_\_\_\_ (2006). "Deriving the Formula of Humanity". In *Groundwork for the Metaphysics of Morals*. Eds. Christoph Horn and Dieter Schönecker. De Gruyter, 2006.
- Korsgaard, Christine (1983). "Two Distinctions in Goodness". In *Creating the Kingdom of Ends*. Cambridge, 1996.
- \_\_\_\_\_ (1986a). "Kant's Formula of Humanity". In *Creating the Kingdom of Ends*. Cambridge, 1996.
- \_\_\_\_\_ (1986b). "Aristotle and Kant on the Source of Value". In *Creating the Kingdom of Ends*. Cambridge, 1996.
- \_\_\_\_\_ (1996a). *The Sources of Normativity*. Cambridge, 1996.
- \_\_\_\_\_ (1996b). *Creating the Kingdom of Ends*. Cambridge, 1996.
- \_\_\_\_\_ (1996c). "An Introduction to the Ethical, Political, and Religious Thought of Kant". In *Creating the Kingdom of Ends*. Cambridge, 1996.
- \_\_\_\_\_ (2002). "Self-Constitution: Action, Identity, and Integrity". Locke Lectures at Oxford University, 2002.
- \_\_\_\_\_ (2003). "Realism and Constructivism in Twentieth-Century Moral Philosophy". *Journal of Philosophical Research*, 2003.
- \_\_\_\_\_ (2004). "Fellow Creatures: Kantian Ethics and Our Duties to Animals". Tanner Lectures at the University of Michigan, 2004.
- Langton, Rae (2007). "Objective and Unconditioned Value". *The Philosophical Review* 116, 2007.
- Rawls, John (1971). *A Theory of Justice*. Harvard, 1971.
- \_\_\_\_\_ (1980). "Kantian Constructivism in Moral Theory". In *Collected Papers*. Ed. Samuel Freeman. Harvard, 1999.
- Ridge, Michael (2005). "Why Must We Treat Humanity With Respect? Evaluating the Regress Argument". *European Journal of Analytic Philosophy*, 2005.
- Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard, 1998.
- Scheffler, Samuel (1992). *Human Morality*. Oxford, 1992.

- 
- Schmidtz, David (1994). "Choosing Ends". *Ethics* 104, 1994.
- Schneewind, J. B. (1991). "Natural Law, Skepticism, and Methods of Ethics". *Journal of the History of Ideas* 52, 1991.
- Shafer-Landau, Russ (2003). *Moral Realism: A Defense*. Oxford, 2003
- Smith, Michael (1999). "Search for the Source". *The Philosophical Quarterly*, 49:196, 1999.
- Thompson, Michael (2004). "What Is It to Wrong Someone? A Puzzle About Justice". In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Eds. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith. Oxford, 2004.
- Velleman, David (2007). *Practical Reflection*. Center for the Study of Language and Information, 2007.
- Wood, Allen (1998a). "Kant on Duties Regarding Nonrational Nature". *Proceedings of the Aristotelian Society Supplemental* 72, 1998.
- \_\_\_\_\_ (1998b). Review of *Creating the Kingdom of Ends*. *The Philosophical Review*, 1998.
- \_\_\_\_\_ (1999). *Kant's Ethical Thought*. Cambridge, 1999.
- \_\_\_\_\_ (2008). *Kantian Ethics*. Cambridge, 2008.