KANT'S PROOF OF THE FORMULA OF HUMANITY

Adam Cureton

University of Tennessee

Abstract: Kant offers the following argument for the Formula of Humanity: Each rational agent necessarily conceives of her own rational nature as an end in itself and does so on the same grounds as every other rational agent, so all rational agents must conceive of one another's rational nature as an end in itself. As it stands, the argument appears to be question-begging and fallacious. Drawing on resources from the Formula of Universal Law and Kant's claims about the primacy of duties to oneself, I propose a quasi-contractualist reconstruction of this puzzling line of reasoning.

In the second section of the *Groundwork for the Metaphysics of Morals*, Kant offers what his admirers and critics alike have described as a "mysterious", "tedious", "obscure", "terse" and "unsatisfactory" argument for the Humanity Formula of the Categorical Imperative (FH):[1]

> The ground of this principle is: *Rational nature exists as an end in itself*. This is the way in which a human being necessarily conceives his own existence, and it is therefore a *subjective* principle of human actions. But it is also the way in which every other rational being conceives his existence, on the same rational ground which holds also for me;* hence it is at the same time an *objective* principle from which, since it is a supreme practical ground, it must be possible to derive all laws of the will. The practical imperative will therefore be the following: *Act in such a way that you treat humanity, whether in your own person or in any other person, always at the same time as an end, never merely as a means.*

[1] Richard Dean (2006: 129) calls the argument "mysterious", John Rawls (2000: 196) calls it "tedious", H.J. Paton (1948: 176) calls it "obscure", Allen Wood (2008: 89) calls it "terse", and Pepita Haezrahi (1962) calls it "unsatisfactory"

Kant adds this "curious" footnote:[2]

* This proposition I put forward here as a postulate. The grounds for it will be found in the final chapter (G, 4.428-9).[3]

We can initially represent Kant's argument for FH as follows:[4]

Premise - Each rational agent necessarily conceives of her own rational nature as an end in itself and does so on the same grounds as every other rational agent.

Conclusion – Therefore, all rational agents must conceive of one another's rational nature as an end in itself.

As it stands, the argument appears to be question-begging and fallacious. Why think that *all* rational persons regard themselves as ends in themselves, and for the same reasons? And even if we all value *ourselves* in this way on the same grounds, why would it follow that we must value *others* in that same way as well? Several strategies have been proposed for cleaning-up Kant's argument, but each has its costs. Drawing on resources from the Formula of Universal Law and what Kant says in the *Metaphysics of Morals* about the primacy of duties to oneself, my aim is to offer a different reconstruction of Kant's reasoning that avoids some main pitfalls and has some advantages over its competitors. If

[2] Wood (2008: 94) also calls Kant's footnote "curious."
[3] I will abbreviate Kant's works by an letter followed by the Academy volume and page numbers: G – *Groundwork for the Metaphysics of Morals* (2002); CPrR – Critique of Practical Reason (2007). MM – Metaphysics of Morals (1996); C – Collins' lecture notes (2001b); and V – Vigilantius' lecture notes (2001a).
[4] Sometimes commentators represent Kant's proof as having two premises, the first saying that a rational agent necessarily conceives of his rational nature as an end in itself, and the second saying that all rational agents conceive of themselves as ends in themselves and do so for the same reasons. According to the other views I discuss, this first premise is either unnecessary or redundant.

successful, the argument as I reconstruct it has a quasi-contractualist character that may be of interest to contemporary followers of Kant who are looking for a way to justify, rather than assume, the unconditional and objective value of rational nature.

1. Background

Kant's argument for the Formula of Humanity plays a prominent role in the larger project he undertakes in his *Groundwork for the Metaphysics of Morals*. His stated aims there are to "seek out" the supreme principle of morality, by arguing in the first two chapters that the various formulations of the Categorical Imperative are implicit in ordinary moral notions of good will and duty, and to "establish" that principle, by showing in the third chapter that our common belief in morality is not illusory (G, 4.392). Kant's striking claim in Chapter Two is that what it is to be subject to duty is to be under a Categorical Imperative, which he defines as an unconditional, comprehensive and necessary requirement of reason addressed to beings who can but might not follow them (G, 4.413-14; 419-21). His initial statement of the supreme moral principle is the Formula of Universal Law (FUL): "'Act only on that maxim by which you can at the same time will that it should become a universal law" (G, 4,421). He goes on to argue, however, that if there is to be a Categorical Imperative, there must be an 'objective end' or 'end in itself' that provides good and sufficient reasons for each and every rational being to follow it (G, 4:427-8). In two quick arguments, Kant contends that rational agents are just such ends in themselves, which leads him to concludes that if there is a Categorical Imperative then it can be expressed as the Formula of Humanity: "*Act in such a way that you treat humanity, whether in your person or in any other person, always at the same time as an end, never merely as a means*" (G, 4.429, Kant's italics).[5]

Before we examine the two arguments Kant gives for FH in more detail, I note a few basic points. First, there is widespread disagreement about what our humanity or rational nature consists in – some

[5] Kant goes on to argue that the Categorical Imperative can also be represented as the Principle of Autonomy "'the Idea of *the will of every rational being as a will that legislates universal law*" (G, 4.431). He leaves open at the end of this chapter the possibility that the idea of a Categorical Imperative may be an illusion, leaving *Groundwork* 3 to argue that morality is not illusory by showing that we really have autonomy of the will of the sort presupposed by our common notion of duty.

think it is our capacity to set ends, others our autonomy, still others our capacity for moral deliberation and action.[6] Second, Kant's commentators diverge over whether the Humanity Formula is essentially the same as the Formula of Universal Law – some think FH is equivalent to FUL by prohibiting maxims that others could not possibly share while others contend that FH requires us to respect, honor and cherish a special dignity or status had by all rational agents.[7] Third, the crucial clause in the Humanity Formula is the one enjoining us to treat humanity in ourselves and others as an end in itself, which includes, and is more comprehensive than, the other clause forbidding us from treating humanity as a mere means. Finally, there are longstanding puzzles about how to apply FH to real-world moral problems without generating moral dilemmas.[8]

Kant's first argument that rational nature in persons is an end in itself is evidently an argument by elimination (G, 4.428). The objects of our inclinations, according to Kant, do not always provide overriding reasons for us to act so they do not have the unconditional and objective worth that would be necessary to ground the Categorical Imperative, which commands us to act whether we are inclined to or not. Similarly, Kant claims that our inclinations themselves are not objectively valuable in the way that would be needed to ground the Categorical Imperative. Non-rational "things" are not ends in themselves either, according to Kant, because they do not give us unconditional and overriding reasons to act. The only remaining candidate is apparently rational nature in persons, so Kant concludes that humanity in ourselves and others is an end in itself.

His second argument, which seems the stronger of the two, is in some ways analogous to Mill's "proof" of the principle of utility, so I will refer to it as Kant's proof of the Humanity Formulation of the Categorical Imperative, or "Kant's proof" for short. Mill argues that because we each desire our own

[6] Christine Korsgaard (1996a) and Allen Wood (1999); (2008), for example, think our rational nature is the power to set ends, Jens Timmermann (2006b) claims that it is the capacity to set moral ends; Richard Dean (2006) argues that having humanity is having a firm and expressed commitment to act morally; Barbara Herman (1993) and Onora O'Neill (1990b) claim that our rational nature is our capacity for moral deliberation and action; and Thomas E. Hill (1992) suggests that humanity is a broader set of rational capacities and dispositions of theoretical, prudential and moral reason.

[7] At one end of this spectrum are O'Neill (1990a), Engstrom (2009) and Sensen (2009), on the other end are Wood (1999); (2008) while Hill (1992), and Korsgaard (1996a) seems to fall somewhere in the middle of these extremes.

[8] Alan Donagan (1977) and Thomas E. Hill (2000); (2003) have been admirably working on this project.

happiness, our happiness is good to each of us, so the general happiness is good to the aggregate of everyone.[9] Similarly, Kant contends that because we each value our own rational nature on the same grounds as everyone else values theirs, we must value the rational nature of all people, so our humanity is an end in itself that must always be treated as such (G, 4.429). Let's look more closely at this argument.

2. The Premise

Most of the discussion of Kant's proof has been focused on:

Premise – Each rational agent necessarily conceives of her own rational nature as an end in itself and does so on the same grounds as every other rational agent.

Why does Kant think that we all necessarily regard ourselves in this way, and for the same reasons?

(P1) One possibility is that we see our rational nature as an end in itself because our non-rational inclinations lead us to do so.[10] We may derive great pleasure and happiness from seeing our rational nature as an objective end, and thinking of our rational nature merely as a thing or tool may cause us great pain and unhappiness, so it is 'necessary' (G, 4.429) that we see ourselves as objective ends in the sense that our empirical psychological makeup inevitably moves us to do so. This would also be a 'subjective' (G, 4.429) principle in the sense that it is a principle that is (1) acted on by a person and (2) based on inclinations. As an empirical generalization, however, this claim is doubtful and anyway it cannot play

[9] (Mill and Warnock 2003: 210) There are notorious problems with Mill's argument, not the least of which is the move from desiring my own happiness to it being good for me and, as we will see, the move from my happiness being good to me to the general happiness being good to the aggregate of everyone. Others have noted the similarities between these arguments in Mill and Kant (Sayre-McCord 2001; Paton 1948; Wood 1999; 2008; Hill 2002a; Haezrahi 1962)

[10] Adrienne Martin (2006: 102, 114) endorses this proposal but looks to Kant's footnote and *Groundwork* 3 for further *rational* grounds for conceiving of ourselves in this way. Paul Guyer (2000: 162-163) advocates for this view as well but thinks Kant's proof is an utter failure.

the requisite role in Kant's proof, as we will see, so we have some reason to look for other interpretations of the Premise.

(P2) A second way of understanding the Premise is that rational agents necessarily conceive of their own rational nature as an end in itself on the grounds that rational nature *per se* is an end in itself.[11] If rational nature is an objective, intrinsic, and agent-neutral value then perhaps it is part of being a rational agent that one see oneself as possessing that value. The point of Kant's proof, on this view, is not to derive the value of humanity from the way we regard ourselves; rather, the "argument" is meant to *convince* people of what is already true, namely that rational nature in general is an end in itself. This interpretation, however, does not fit very well with Kant's use of inferential vocabulary in the relevant passage, which seems to suggest he is giving some sort of argument for FH rather than merely helping people see what is in fact valuable if only they could move beyond their parochial concerns. A further difficulty with any reading of Kant that assumes an antecedent intrinsic value is Kant's well-known claims about the "right" being prior to the "good" such as: "*the concept of good and evil must not be determined before the moral law (for which, as it would seem, this concept would have to be made the basis) but only (as was done here) after it and by means of it*" (CPrR, 5.63, Kant's italics).[12]

(P3) The most influential account of why Kant thinks rational agents necessarily view themselves as objective ends is the clever and vexing 'regress argument' of Christine Korsgaard and Allen Wood.[13] They interpret Kant as follows: When we set ends, which we must do in order for there to be rational action and a Categorical Imperative, we take those ends to be objectively good in the sense that all rational agents have reasons to value our own ends just because we set them. We necessarily regard our ends to be objectively good and take the source of their value to be our endorsement of them, so it follows

[11] (Sayre-McCord 2001; Wood 1999; 2008)

[12] Allen Wood, who has developed an impressive, subtle and sophisticated interpretation of Kant's ethical theory that regards rational nature as an ungrounded, intrinsic and agent-neutral value, is well aware of this and other passages in which Kant seems to prioritize the "right" over the "good" (2008: 155). My aim here is not to criticize this view but rather to see if there is a different interpretation of Kant's Proof those who agree that rational nature is a thick and substantive value but think that it must somehow be grounded in the Categorical Imperative rather than the other way around.

[13] (Korsgaard 1996a; Wood 1999; 2008) For discussion and criticism see (Arroyo 2011; Dean 2006; Denis 2007; Martin 2006; Bruton 2000; Kerstein 2009; Schneewind 1998)

that we must regard our capacity to set ends (i.e. our rational nature) as unconditionally and objectively good and so an end in itself.

One lingering problem with this argument is whether we really do attribute objective value to the non-rational ends we set or whether instead we see our attaining our ends as good *for us* without yet presupposing that we automatically generate reasons for others to respond favorably to our personal plans and projects merely by forming them – having a project of finishing Robert Caro's multi-part biography of Lyndon Johnson may generate certain reasons *for me* to do so but if we do not add in an additional moral principle of, say, beneficence then it is difficult to see why all rational agents have reason to help me get to reading or see my finishing the books as good *tout court* as opposed to good *for me*.[14] A further issue is about why we must regard our capacity to set ends as objectively and unconditionally valuable even if it is the source and condition of our objectively valuable ends – Elvis may have the power to make things objectively good by touching them, but it does not seem to follow that this capacity itself is objectively good *in itself* even if its deliverances are good in virtue of his touch. It may be that having and exercising that power to make things good by handling them is objectively good *as a means* to bringing about good things, but it is far from clear why this power would be good in itself, as the regress argument requires.[15] Finally, Korsgaard and Wood disagree about whether the regress argument is supposed to *establish* that someone's rational nature is objectively and unconditionally valuable or whether setting ends merely commits a rational person to *regarding* or *seeing* her rational nature as having that sort of value, which it has whether not she recognizes it or not. In other words, does setting ends make that capacity an end in itself or is it already an end in itself and rational reflection in the form

[14] Agent-relative reasons, as they are sometimes called, are reasons with an ineliminable back-reference to the agent who has the reason – that *my* brother needs money may be a reason for *me* to give him a loan but it may not be a reason for others to do so – whereas an agent-neutral reason is a reason with no such back-reference – that someone is in pain is a reason for anyone to help them (Petit 1987; Parfit 1984; Nagel 1978) There are certainly texts that seem to support the claim that in setting ends we take them to generate agent-neutral reasons (e.g. G, 4,412-14; 423; 430, MM 6:393, CPrR 5:74) but my reconstruction of Kant's proof involves a different way of understanding the role of agent-relative reasons in Kant's ethical theory can fit these passages as well.

[15] This case is generalized from an example of Geoff Sayre-McCord's in which some pink Cadillac at Graceland is extrinsically valuable because Elvis touched it.

of the regress argument reveals this value to us?[16] If the latter then P3 is just a more sophisticated version of P2.

(P4) A fourth way of understanding the Premise is that it is simply part of our rational nature to regard our humanity as an end in itself because we have autonomy of the will.[17] Kant explains in the footnote to the passage where he gives his proof of FH that the reason we take our rational nature to be an end in itself is given in *Groundwork* 3, where he argues that rational agents by necessity acknowledge that they have autonomy of the will. A rational person "necessarily conceives his own existence" (G, 4.429) in this way, as an end in itself, and she does so, according to this interpretation, on the same grounds as other rational agents, namely that he possesses autonomy of the will. If we ask why it is a necessary part of being a rational person that one conceive of oneself in this way and on these grounds, we may find that Kant did not attempt to argue directly for the Premise or explicitly explain why he thinks it is true, choosing instead to leave its justification to ordinary moral understandings about the nature of a rational and reasonable person and perhaps also to the Fact of Reason (CPrR, 5.151-2; 155-6). Kant's view may have been that once we understand the Premise in this way, no further justification is needed for it, although some of us may not be satisfied with stopping moral inquiry here.[18]

[16] Allen Wood's (2008: 94-96) position on this issue is complicated. Wood does not think we "confer" value on things, as Korsgaard suggests, but does think that rational nature is the sole "source" of objective value. If rational nature is the only objective, intrinsic and agent-neutral good then other things are also good in virtue of their relation and contribution to this original good, things like food, shelter, education, etc., which are then good whether or not we actually set them as ends. Kant's proof and the regress argument, as Wood sees it, does not *establish* or *demonstrate* that rational nature is an end in itself, as it is meant to for Korsgaard; it merely shows us that in setting ends we presuppose an objective value that is *already there*, which presumably means that even if our practices were different and setting non-rational ends did not involve ascribing them objective value, our rational nature would still be an end in itself. If this is correct, Wood's position is closely akin to the third way of justifying the Premise.

[17] (Rawls and Herman 2000: 196-199; Hill 1992; 2002a)

[18] Martin attempts to combine (2) and (3) by suggesting that every time we act, we conceive of ourselves as having and acting from a sort of autonomy that we see as an ultimate and unconditional end, which, according to her, implies that "to be autonomous is (*inter alia*) to be an end in itself" (Martin 2006: 116). Her view is interesting in its own right, but she is explicit that her conception of autonomy is not Kant's, so as it stands her view is not a suitable enough interpretation of Kant's proof. One might also worry that if autonomy is understood as identification with one's core values then seemingly genuine actions involving weakness of will are either block her regress argument or they are implausibly excluded from being actions at all. A further concern with her argument is whether autonomy in this sense is an unconditional end that we act for in all circumstances – sometimes, it seems, we should give up our core values if they are immoral or if acting on them is immoral.

3. The inference

Kant's inference from the Premise to his conclusion is just as problematic:

Conclusion – Therefore, all rational agents must conceive of one another's rational nature as an end in itself.

How (if at all) might we explain the move from each of the four interpretations of the Premise we just discussed to the conclusion that we are all committed to regarding rational nature in others as an end in itself?

(C1) Even if we assume that all rational agents unavoidably conceive of themselves as ends in themselves because this makes them happy, it does not logically follow that they must regard one another as ends in themselves. Just because we may *desire* the same thing, say happiness, this does not give us any *reason* to satisfy that desire, according to Kant, which means that even if conceiving of one another as ends in ourselves contributes to general happiness, this alone could not make us ends in ourselves (G, 4.431). Moreover, the Conclusion is supposed to be a universal principle, which Kant thinks cannot be established on the basis of empirical generalizations (G, 4.431).[19] In order for Kant's proof to succeed, his Premise must say that we are *rationally* required to conceive of ourselves as ends in ourselves, rather than that we do so merely on the basis of empirical inclinations.

(C2) Suppose next that we assume that rational agents necessarily conceive of their own rational nature as an end in itself on the grounds that rational nature *per se* is an end in itself. Because rational nature is an intrinsic, agent-neutral and objective value, we have reasons to regard it in ourselves and others as such. The role of Kant's "proof" would then be to elucidate and reveal this objective value to rational persons who may be biased and distracted by the splendor of their own rational nature and need reminding that the rational natures of others are objective ends as well. Some Kantians are reluctant to interpret Kant as appealing to a value of this sort as a starting-point, particularly in a context in which

[19] (Pogge 1998: 197; Hill 2002a: 124)

Kant seems to be giving an argument for the existence of such a value. Kant's inferential language of 'therefore' and 'hence' (G, 4.431) seems to suggest he is presenting an argument of that sort rather than helping us engage in rational reflection in which we "cite what we do, and what we must represent ourselves as thinking and doing, when we form preferences, set ends, and make decisions, and then to argue that these actions, thoughts, and representations are best understood as recognizing something as an ultimate value."[20]

(C3) We can consider the last two versions of the Premise (P3 and P4) together because the same problem arises when we try to derive the Conclusion from either of them. Suppose it is part of our rational nature to conceive of *our own* rational nature as an end in itself on the basis of our capacity to set ends or our autonomy of the will. Kant's proof could then be understood as addressed to any rational agent, pointing out to him that other rational agents have the rational nature that, as he sees it, makes him an end in himself, and there is no relevant difference between his rational nature and those of others, so he must conceive of their rational nature as an end in itself as well. If this argument succeeded, it would show that it is a necessary feature of rational agency to conceive of humanity as an end in itself without appealing to an already existing, agent-neutral value.

This argument is not air-tight, however, because of the unlikely but conceivable case in which a rational agent conceives of himself as an end in himself because of *his own* rational nature, he sees the fact that his rational nature as *his* as a relevant difference between it and others, and wonders what reason *he* has to regard humanity in general as an end in itself. Call this the egoistic fallacy. In order for Kant's proof to establish the conclusion that all rational agents must conceive of rational nature in this way, we need rational grounds for incorporating this possibility into the argument or setting it aside.

In response, one might look back to the 'regress argument' and claim that because our egoist has the capacity to set objective ends, he must regard his rational nature as an end in itself that provides reasons to all rational agents; others have that same power as well, so their rational nature provides reasons for him. This does not avoid the egoistic fallacy, however. Korsgaard says that regarding just

[20] (Wood 2008: 90) See also (Wood 1999; Haezrahi 1962; Sayre-McCord 2001)

one's own *happiness* as unconditionally good is a "remarkable feat of egocentrism" but it would be even more remarkable, yet still conceivable, for someone to regard *his* power to set ends, and *his* power alone, as objectively and unconditionally good.[21] This would mean that, as he sees it, others have reasons to treat his rational nature as as an end in itself but he sees no reason to admit that that they are ends in themselves as well, for according to him they lack the power, which only he possesses, to set objective ends.

Another way to try to avoid the egoistic fallacy is to appeal to common moral understandings in an attempt to set aside the outlier case of someone who genuinely regards his rational nature as an end in itself because it is *his*. Such a person is clearly morally defective, so he may not be the kind of person we have to convince that humanity in general is an end in itself – we may even wonder whether a rational person could sincerely have and maintain such an attitude. It may be enough for Kant's purposes just to show reasonable, conscientious people that they have good and sufficient reasons to treat humanity as an end in itself. In the same vein, Korsgaard and Dean draw on some vague suggestions Kant makes in the *Critique of Practical Reason* that "a universal law of nature makes everything harmonious," so if the Categorical Imperative merely required everyone to regard her own rational nature as an end in itself, but not others, then the "most extreme opposite of harmony would follow, the worst conflict" because "the will of all has not one and the same object but each has his own" (CPrR, 5.28).[22] My reconstruction of Kant's proof appeals to some of these same ideas, but it is worth noting that the context of this passage is very different from Kant's arguments for the Formula of Humanity. In the Second Critique passages, Kant is focused on how a universal principle of pursuing one's own *happiness* would not be harmonious because we will inevitably want the same things, but it is less clear that a world of egoists who regard just themselves as ends in themselves would necessarily be in moral conflict with one another on that account – their desires and inclinations may pit them against one another, but they do not think they have moral reasons to pursue their own happiness, so it is possible that they will just be morally indifferent to one

[21] (Korsgaard 1996a: 122)
[22] (Korsgaard 1996a: 122; Dean 2006: 128-129)

another, which would be a kind of harmony. One wonders, moreover, why Kant thinks moral principles must be harmonious in the sense that in following them we have the same object rather than our various proprietary ones. A moral principle allowing everyone to pursue her own happiness would, according to Kant, result in "complete annihilation of the maxim itself and its purpose," which suggests that disharmony is a matter of failing to satisfy the Formula of Universal Law, but it is unclear whether a principle only requiring people to regard themselves as ends in themselves would similarly fail that test. Kant could also be appealing to a sort of harmony found in natural teleology, but there is deep disagreement about whether Kant's views about natural teleology play a foundational role in his moral theory. Lastly, proponents of the 'regress argument' cannot dismiss the radical egoist I described on these grounds because, as our egoist sees it, no one besides himself has the power to set objective ends, only he has that special power, so there is no potential for disharmony between his objective ends and those of others.

Commentators of Kant's proof are well aware of the egoist fallacy, so even when they do not discuss in detail what to say about an egoist who conceives of his rational nature as an end in itself for reasons that contain essential reference to himself, most of them assume there must be resources in Kant for handling this possibility.[23] The case of such an extreme egoist is so remote and far-fetched that perhaps it is safe to assume that Kant has enough resources that we need not be side-tracked by it.

Where does this discussion of Kant's proof leave us? Having surveyed the main contenders for interpreting Kant's proof and noted some problems with each, some of us may find ourselves in the following position. We would like, if at all possible, to find an interpretation of his argument for FH that (1) does not appeal to empirical premises about what we happen to desire, (2) does not assume that rational nature is already an intrinsic, agent-neutral and objective value of humanity but instead (3) attempts to establish that as a conclusion, and also (4) avoids the fallacy of inferring that everyone must regard rational nature as an end in itself because all rational agents conceive of their own rational nature in that way, for some of them may do so because it is *theirs.* Others may wish to pursue one of these

[23] (Rawls and Herman 2000: 197; Pogge 1998)

other strategies, and perhaps these are our only options available (5) without treating the proof as an utter failure, but I now wish to suggest an alternative interpretation that satisfies these five desiderata and, surprisingly, points toward a non-traditional way of understanding the Formula of Universal Law.

4. Rational self-regard and regard for others

According to the interpretation of Kant's proof I propose, what is widely regarded as a more or less trivial case of someone who treats her own rational nature as an end in itself because it is *hers* actually lies at the heart of what Kant intends that argument to do, which is to explain why someone who has rational regard *for herself* should regard humanity in *others* as an end in itself as well. Kant's commentators have puzzled over his repeated insistence that duties to oneself are the foundation of duties to others (MM, 6.418; C 27.341; V 27 579-80)[24], but one way of understanding his point is that his normative ethical theory *begins* with an idea of a rational person who conceives of her own rational nature as an end in itself because it is hers. Such a person recognizes agent-relative reasons to treat herself accordingly but does not think it immediately follows from the way she regards herself that others thereby have reasons to treat her rational nature as an end in itself as well. Rational agents, however, are also committed to a kind of moral *reciprocity* and *mutuality* in the form of the Formula of Universal Law. In the spirit of moral contractualist theories, each one of them is willing to act on principles that are, in a sense, acceptable to all, so rational agents could agree to treating humanity in oneself and others as an end in itself because this principle allows each person to have proper regard for herself while affording the same to others.

More specifically, Kant starts the relevant passage with the claim that a representative rational agent, say P, necessarily conceives of his own rational nature as an end in itself. Doing so is *necessary* for P, not as a matter of his empirical psychology, but in the sense that it is *rationally* necessary for P to regard his own humanity as an end in itself, which means that P is rationally committed and disposed to conceive of his own rational nature as an end in itself and if he were fully rational he would succeed at

[24] (Timmermann 2006a; Reath 1998; Potter 2002; Denis 1997)

always treating himself accordingly. Kant describes this way of regarding himself as a *subjective* principle of human action in order to indicate that the principle is not only (1) a principle on which P acts but is also (2) *about P himself*.

Kant goes on to claim that every other rational being conceives of his or her own rational nature as an end in itself. I have followed the common practice of formulating this part of Kant's proof in terms of the way *all* rational agents conceive of their humanity, but we can ask why Kant begins the passage by describing how a *particular* rational agent conceives of himself? He does so in order to signal that the argument is addressed to the subjective perspective of a person who conceives of his own rational nature as an end in itself but wonders why, if at all, he should view others in that way as well. If Kant can give P sufficient reasons to have proper regard for humanity in others then he will have explained to each one of us, from our own standpoints, why we should do so as well, which is just the approach we would expect from a contractualist style of argument.

We are told that each of us regards our own rational nature as an end in itself "on the same rational ground," and a footnote tells us that these reasons for conceiving of ourselves in this way will be more fully described in the final chapter. Kant is evidently referring to the argument he gives there that all rational agents have autonomy of the will (G, 4.446-8). His line of reasoning is notoriously difficult, but its main structure is that rational agents must take themselves and others to be free in the negative sense that their wills can do things "independently of alien causes determining it" (G, 4.446), which entails that they have autonomy of the will in the positive sense that they have the rational capacity and disposition to act on unconditionally rational principles.

Rational agents must act "*under the Idea of freedom*" (G, 4.448) by taking ourselves to have the ability to make things happen without being determined by causes that are independent of our own choices, commitments, plans and judgments. Seeing ourselves in this way is not a contingent psychological feature of us but is rather implicit in our taking up a first-person, deliberative standpoint, which we must do as rational agents. A rational agent must not only take himself to be free in this sense when deliberating about what to do, but Kant claims that he must also see other rational agents as

negatively free as well in order to understand them as choosing, intending and deciding for reasons. Because having negative freedom supposedly entails having autonomy of the will, Kant thinks each of us has sufficient reason to attribute autonomy of the will to ourselves and every other rational being.

One tempting way of reading this discussion of autonomy back into Kant's proof of FH is that each of us regards our own rational nature as an end in itself because we take ourselves to have autonomy of the will, but we must also regard others as autonomous as well, so we all have sufficient reason to regard one another's rational nature as an end in itself because we each have an autonomous will. We have already reviewed some of the problems with this basic strategy, however, most notably that a rational agent may see herself and others as having autonomy of the will but choose to regard only her own rational nature as an end in itself because she regards her own autonomous will as special. Her will is hers, in her view, it is part of who she is and she is in charge of it and not others.

According to my proposal, all persons conceive of their own humanity as an end in itself and they do so on the basis of the same aspect of themselves, namely their *own* autonomous will – I regard myself as an objective end in virtue of my autonomous will, you do the same because of yours. By emphasizing that rational agents necessarily regard one another as having autonomy of the will, Kant shows that his proof of FH is not addressed to a radical skeptic who denies that there are other rational agents besides himself or to a radical egoist who only cares about her own interests. The proof is instead meant to convince a rationally self-regarding and autonomous person who sees herself as living in a world of like agents who she is committed to reciprocating with but wonders why she should regard their rational natures in the same way that she regards her own.

How might we persuade a person in this position that she should conceive of others' humanity as an end in itself? My suggestion is that in his proof of FH Kant is implicitly appealing to our most fundamental rational commitment as an autonomous agent, which he describes in his *Groundwork* 3 discussion of autonomy as the "formula of the Categorical Imperative" to act "on no other maxim than

one that can also have [itself as] a universal law for its object" (G, 4.447).[25] This is sometimes called the Principle of Autonomy (PA). Having autonomy of the will, Kant explains, requires us to see ourselves as subject only to universal laws that we and others legislate. Moreover, prior to his proof of FH, Kant has argued that the common notion of duty, which all rational agents accept, presupposes the Formula of Universal Law (FUL) – "Act only on that maxim by which you can at the same time will that it should become a universal law" (G, 4,421). Both of these principles are therefore available and promising resources for understanding Kant's proof of FH.

Kant claims that PA and FUL are somehow equivalent (G 4.436), and interpretations of both formulations differ, but one common problem between them is to specify the standards that determine whether or not someone can legislate or will a maxim as a universal law. These standards are not merely a matter of what people happen to want or contingent factors about their psychology but instead specify reasons for what we can and cannot rationally legislate or will. Various standards have been proposed: a rational agent wills the "normal and predictable results" of his actions, he avoids undermining his own purposes, he takes the necessary means to his ends or gives up his ends, or wills the capacity for effective willing.[26] When deciding whether he can rationally legislate or will his proposed maxim as a universal law, a rational agent looks for an inconsistency between his maxim, along with its typical and foreseeable results, underlying purposes, or any other necessary means it may involve, and everyone acting or being permitted to act in that way.[27] Commentators disagree about whether these views fit the text, whether some of them are actually standards that are inherent in practical reason, and whether they can generate all of the moral duties that seem evidently correct.

Thomas E. Hill has proposed that standards of rational willing include, along with formal ones of consistency and coherence, substantive requirements about what ends a rational person must will.[28] He argues that these substantive standards should be imported from other formulations of the Categorical

[25] This translation is slightly revised from Hill and Zweig's.
[26] (O'Neill 1975: 63-93; Korsgaard 1996b; O'Neill 1990a; Herman 1993: 121-122)
[27] For discussion about whether or not FUL is a principle of requirement or permissibility see (Pogge 1998) who cites unpublished work of Thomas M. Scanlon.
[28] (Hill 2002b) See also (Guyer 1995).

Imperative, particularly FH. If rational agents necessarily treat humanity in themselves and others as an end in itself then any proposed maxim that, when universalized, would be incompatible with regarding humanity in this way cannot be legislated or willed as a universal law by any rational agent. Hill supplements FUL with Kant's claim that rational agents are rationally committed to treating humanity as an end in itself, but our concern has been whether Kant has an argument that rational agents must regard one another in this way.

A different strategy for incorporating substantive requirements of rational willing is to interpret FUL and PA as principles of moral contractualism that basically requires us to act only in ways that are justifiable to everyone, where the rational standards of acceptability are either formal or substantive and *self-regarding*. More specifically, P should not act on a maxim that, if universally adopted, could not be rationally willed by each and every rational agent in light of what they each rationally will *for themselves*. If the universal counterpart of P's maxim conflicted with the rational self-regard of another person then that maxim is impermissible. Rational agents are committed to their own rational interests but they are also willing to reciprocate with other rational agents in the sense that they will constrain themselves by principles that are acceptable to everyone.

What, then, are the self-regarding and substantive standards of rational willing that determine whether a maxim can be legislated or willed as universal law? As I see it, this is one of the questions that Kant's discussion of FH is meant to answer and helps to explain its role in the overall argument of *Groundwork* 2. Perhaps the most important standard of rational willing, I have been suggesting, is that rational agents conceive of their own rational nature as an end in itself. Regarding oneself in this way not only involves acting accordingly toward oneself but also standing up for oneself with others, insisting that they treat one's humanity as an end in itself and refusing to allow oneself to be treated as a relative end. A rational agent is also autonomous and so is committed to acting only on maxims that are acceptable to all rational agents in virtue of their self-regarding rational standards, and those standards prominently include treating one's own rational nature as an end in itself. Any principle that permitted the rational nature of someone to be treated as anything other than an end in itself is inconsistent with the self-

regarding rational commitments of such a person – no one could rationally accept a moral principle that permitted others to treat her as a relative end. Each rational agent could rationally agree, therefore, to a principle that says that rational nature in oneself and others must be treated as an end in itself, for this principle is consistent with each person having proper regard for herself. Rational agents, in other words, are rationally disposed to ensure that others treat her as an end in herself, but each one is also willing to act on principles that are, in a sense, acceptable to all, so rational agents must treat humanity in one another as an end in itself. That, as I see it, is how Kant's proof of FH proceeds.

More formally, we can express the argument as follows:

(Premise 1) Each rational agent necessarily conceives of her own rational nature as an end in itself and does so because of her own autonomous will.

(Premise 2) Rational agents are necessarily committed to acting only on maxims that can be legislated or willed as universal law by self-regarding rational agents.

(Conclusion) Therefore, because no rational agent can rationally will a universal law that allows others to treat her as anything other than an end in herself, all rational agents must conceive of one another's rational nature as an end in itself.

This argument satisfies the four desiderata I defined. (1) It avoids empirical premises about what we contingently want and (2) assumptions about antecedent and agent-neutral values. (3) My reconstruction attempts to establish as a conclusion that all rational agents have reasons to treat one another's rational nature as an end in itself without (4) committing the egoistic fallacy while (5) showing Kant's proof of FH to be forceful and philosophically interesting.

Conclusion

What this argument establishes, if it is successful, is a "thin" version of FH, one that is essentially equivalent to FUL but emphasizes that we are to treat others in ways that they can rationally share. The principle does not yet tell us much about what, in particular, it takes to treat rational nature in oneself and others as an end in itself. Yet in the *Metaphysics of Morals* Kant seems to invoke a "thick" version of FH, one that has specific implications for how to treat humanity in oneself and others that go beyond the basic contractualist requirement to treat others in ways that are compatible with their rational self-regard. Some commentators handle this apparent gap by deflating the specific duties that are supposed to derive from FH (or its near equivalent, the Supreme Principe of the Doctrine of Virtue) or else reading a thick version of FH back into the *Groundwork* in a way that obscures how it can be equivalent to FUL.

The argumentative strategy I have proposed offers a unified account of these matters because the generality of its conclusions can vary depending on how specifically the self-regarding standards of rational willing are stated. In the *Groundwork*, where Kant tries to seek out and justify the supreme principle of morality, he presents FH as a thin principle that derives from FUL and a very general rational requirement to treat one's rational nature as an end in itself, without filling in the details about how, in particular, to do so. As I argue elsewhere, in the *Metaphysics of Morals*, we find him using the same strategy by clarifying what is involved in treating one's own rational nature as an end in itself and then using FUL to derive duties of beneficence and respect for others.[29] Kant claims that we are rationally required to pursue our own happiness (although he emphasizes that we have no duty to do so) and we are rationally required to maintain our self-respect and avoid servility. As embodied rational agents, we must therefore will that others sometimes help us to achieve our ends and refrain from undermining our self-respect, because these are rationally necessary in order for us to achieve our rationally required and self-regarding ends. Each of us is also committed to FUL, so we can will universal laws requiring beneficence and respect for others. Specific other-regarding duties, according to Kant, are those that can be willed by all rational agents in virtue of what they necessarily will as self-regarding agents.

[29] [Citation suppressed for blind review]

Works cited

Arroyo C. (2011) Freedom and the source of value: Korsgaard and wood on Kant's formula of humanity. *Metaphilosophy* 42: 353-359.

Bruton SV. (2000) Establishing Kant's formula of humanity. *Southwest Philosophy Review* 16: 41-49.

Dean R. (2006) *The Value of Humanity in Kant's Moral Theory,* Oxford: Oxford University Press.

Denis L. (1997) Kant's ethics and duties to oneself. *Pacific Philosophical Quarterly* 78: 321–348.

Denis L. (2007) Kant's formula of the end in itself: Some recent debates. *Philosophy Compass* 2: 244–257.

Donagan A. (1977) *The Theory of Morality,* Chicago: University of Chicago Press.

Engstrom S. (2009) *The form of practical knowledge,* Cambridge, Mass.: Harvard University Press.

Guyer P. (1995) The possibility of the categorical imperative. *Philosophical Review* 104: 353-385.

Guyer P. (2000) *Kant on freedom, law, and happiness,* Cambridge: Cambridge University Press.

Haezrahi P. (1962) The concept of man as end-in-himself. *Kant-Studien* 53: 209-224.

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

Hill TE. (1992) Humanity as an End in Itself. *Dignity and Practical Reason in Kant's Moral Theory.* Ithaca: Cornell University Press, 38-57.

Hill TE. (2000) Donagan's Kant. *Respect, pluralism, and justice : Kantian perspectives.* Oxford ;New York: Oxford University Press, 119-151.

Hill TE. (2002a) Editor's Introduction. In: Kant I, Hill TE and Zweig A (eds) *Groundwork for the metaphysics of morals.* Oxford ; New York: Oxford University Press, 19-91.

Hill TE. (2002b) Hypothetical Consent in Kantian Constructivism. *Human welfare and moral worth.* Oxford: Oxford University Press, 61-96.

Hill TE. (2003) Treating Criminals as Ends in Themselves. *Annual Review of Law and Ethics* 11: 17-36.

Kant I. (2001a) Kant on the metaphysics of morals: Vigilantius's lecture notes. In: Heath PL and Schneewind JB (eds) *Lectures on ethics.* Cambridge: Cambridge University Press, 249-452.

Kant I. (2001b) Moral philosophy: Collins's Lecture notes. In: Heath PL and Schneewind JB (eds) *Lectures on ethics.* Cambridge: Cambridge University Press, 37-222.

Kant I and Gregor MJ. (1996) *The Metaphysics of Morals,* New York: Cambridge University Press.

Kant I and Gregor MJ. (2007) *Critique of Practical Reason,* Cambridge: Cambridge University Press.

Kant I, Hill TE and Zweig A. (2002) *Groundwork for the Metaphysics of Morals,* Oxford: Oxford University Press.

Kerstein S. (2009) Treating others merely as means. *Utilitas* 21: 163-180.

Korsgaard CM. (1996a) Kant's Formula of Humanity. In: Korsgaard CM (ed) *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press, 106-122.

Korsgaard CM. (1996b) Kant's Formula of Universal Law. In: Korsgaard CM (ed) *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press, 77-105.

Martin A. (2006) How to argue for the value of humanity. *Pacific Philosophical Quarterly* 87: 96-125.

Mill JS and Warnock M. (2003) *Utilitarianism and On liberty,* Malden, MA: Blackwell Pub.

Nagel T. (1978) *The Possibility of Altruism,* Princeton: Princeton University Press.

O'Neill O. (1975) *Acting on principle,* New York: Columbia University Press.

O'Neill O. (1990a) Consistency in Action. *Constructions of Reason.* Cambridge: Cambridge University Press, 81-104.

O'Neill O. (1990b) *Constructions of Reason,* Cambridge: Cambridge University Press.

Parfit D. (1984) *Reasons and Persons,* Oxford: Oxford University Press.

Paton HJ. (1948) *The Categorical Imperative; A Study in Kant's Moral Philosophy,* Chicago: University of Chicago Press.

Petit P. (1987) Universality Without Utilitarianism. *Mind* 72: 74-82.

Pogge T. (1998) The Categorical Imperative. In: Guyer P (ed) *Kant's Groundwork of the metaphysics of morals: critical essays.* Lanham, Md.: Rowman & Littlefield, 189-213.

Potter N. (2002) Duties to Oneself, Motivational Internalism, and Self-Deception in Kant's ethics. In: Timmons M (ed) *Kant's Metaphysics of morals: Interpretative essays.* Oxford: Oxford University Press, 371-390.

Rawls J and Herman B. (2000) *Lectures on the History of Moral Philosophy*, Cambridge: Harvard University Press.

Reath A. (1998) Self-Legislation and Duties to Oneself. *Southern Journal of Philosophy* 36: 103-124.

Sayre-McCord G. (2001) Mill's "Proof" of the Principle of Utility: A More than Half-Hearted Defense. *Social Philosophy and Policy* 18: 330-.

Schneewind JB. (1998) Korsgaard and the unconditional in morality. *Ethics* 109: 36-48.

Sensen O. (2009) Kant's Conception of Human Dignity. *Kant-Studien* 100: 309-331.

Timmermann J. (2006a) Kantian duties to the self, explained and defended. *Philosophy* 81: 505-530.

Timmermann J. (2006b) Value without Regress: Kant's 'Formula of Humanity' Revisited. *European Journal of Philosophy* 14: 69-93.

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

Wood AW. (2008) *Kantian Ethics*, Cambridge: Cambridge University Press.