Freedom, Morality, and the Propensity to Evil

Samuel Kahn
Indiana University - Purdue University Indianapolis

In Book I of the *Religion within the Boundaries of Mere Reason* Kant offers an explanation of freedom and moral good and evil that is different from that offered in the *Groundwork for a Metaphysics of Morals*. My primary goal in this paper is to analyze and elucidate this new theory. My secondary goal is to contrast this new theory with the older one that it is replacing. I argue that the new theory, which centers on the idea that evil involves a sort of misprioritizing, enables Kant to get around two problems associated with the older theory. As will be seen, this has implications for two main debates in the secondary literature: the *wille-willkür* debate and the debate about the exegetical plausibility of the so-called regress argument for the formula of humanity.

Accordingly, my paper will be organized into three main sections. In section one, I will discuss the relationship between freedom of choice and moral good and evil as set out in Book I of the *Religion within the Boundaries of Mere Reason* (R1). In particular, I will focus on the arguments that Kant offers to substantiate his claim that "the human being is by nature evil" (AA RGV 06:32) because I think that this claim is central to R1 as a whole. Focusing on it will facilitate and allow a nucleation point for the discussion in section one. I will transition between sections one and two by looking at Kant's conception of evil and how this allows him to claim at the beginning of R1 that "*considered in themselves* natural inclinations are *good*" (AA RGV 06:58). In section two, I will briefly outline two distinct but related problems that arise from Kant's conception of freedom and morality in the *Groundwork for a Metaphysics of*

Samuel Kahn, Freedom, Morality, and the Propensity to Evil,
KSO 2014: 65-90, Posted April 2, 2014
www.kantstudiesonline.net

Morals. I refer to these as the 'Sidgwick problem' and the 'Dignity and Worth' problem, respectively. After articulating these problems, I will try to show how the theory offered in R1 solves both of them. In discussing the Sidgwick problem, I will set out three distinct positions in the *wille-willkür* debate and explain where I fit in. In discussing the Dignity and Worth problem, I will connect my discussion up with the regress argument for the formula of humanity. Finally, in the third section of the paper, I will conclude with a brief recapitulation, some comments on some of the implications of this new theory, and an open-ended question as to what Kant's claim about the human being's propensity to evil, taken in conjunction with an empirical claim he makes in the introduction to R1, means for his theory of the highest good and the practical postulates, especially as this theory is outlined in the *Critique of Practical Reason*.[1]

Section 1: the human being is by nature evil

In the general introduction to RI Kant states his goal for the succeeding sections: "we shall say… of one of these characters [i.e. the good or the evil]… that it is *innate* in [the human being]" (AA RGV 06:21). However, this already raises a number of questions: what does it mean that the human being should be good (or evil) 'innately'? Kant offers other glosses of his goal here: he intends to establish whether "the human being is by nature good… [or] he is by nature evil" (AA RGV 06:21). But this is useful only insofar as it hopefully will prevent us from getting too bogged down with particular words, like "innate." What Kant means by these turns of phrase still

[1] *The text of the following footnote was added post-publication on 26/04/2014:* In citing material from *The Critique of Pure Reason*, I have used the standard A and B pagination for the 1781 and 1787 editions respectively. All other citations of Immanuel Kant's works are identified by the relevant abbreviation together with the volume and page number of *Kant's gesammelte Schriften*. All translations are taken from the Cambridge Blue Series of Kant's works edited by Paul Guyer and Allen Wood.

remains unsettled. Moreover, it is *prima facie* implausible that this short list of options is exhaustive. Cannot a human being be both good and evil or neither good nor evil? Surely evil comes in degrees: someone who tells the occasional white lie cannot be put on a level with an Adolf Eichmann.

The first of these questions feeds into the second and offers us the first look at the connection between freedom and morality in the *Religion within the Boundaries of Mere Reason*. Kant explains that by the "nature" of a human being he means "the subjective ground… of the exercise of the human being's freedom in general" (AA RGV 06:21).[2] In other words, though a human being can be good or evil by nature, he still "holds within himself a first ground for the adoption of good or evil… maxims" (AA RGV 06:21) because his freedom of choice precludes any "determination through natural causes" (AA RGV 06:21).[3] Indeed, a human being, since he is free, "cannot be determined to any actions through any incentive *except so far as the human being has incorporated it into his maxim*" (AA RGV 06:24). In other words, a human being is free to act on any maxim – the one thing that a human being cannot do is foreswear his/her freedom of choice (in any but a trivial fashion) (AA RGV 06:27). Therefore, any failure to incorporate the moral law into one's maxim cannot be a negative failure (i.e. morally neutral) but must be considered as a positive failure (AA RGV 06:24)[4] – and therefore when considering the human being as an intelligible being, s/he can be judged only either morally good or morally evil, not both (or neither) (AA RGV 06:23-25n).[5] In other words, the options above are, indeed, exhaustive, although (as will be explained below) this does not entail that evil does not come in degrees.[6] I will bring up this rigid dichotomy again in the conclusion.

---

[2] See also AA RGV 06:25.
[3] See also AA RGV 06:25.
[4] For a similar reading, see Henry Allison, "On the very idea of a propensity to evil," *The Journal of Value Inquiry* 2002, 36: 337-348.
[5] See also AA RGV 06:25-26.
[6] See also AA Vorl 27:511.

Kant next presents our three predispositions to the good. The first is animality, "physical or merely *mechanical* self-love" (AA RGV 06:26). The second is humanity, "self-love which is physical and yet *involves comparison*" (AA RGV 06:27). The third is personality, "the susceptibility to respect for the moral law *as of itself sufficient incentive to the power of choice*" (AA RGV 06:27).[7] As predispositions, these three characteristics are innate and cannot be represented as otherwise than innate (AA RGV 06:29). They are "original," meaning that they "belong with necessity to the possibility of this being" (AA RGV 06:28). However, they are still only *incentives* to the free power of choice and, therefore, even with these good predispositions, a good or evil character "is something that can only be acquired" (AA RGV 06:27). In other words, a good character is still only something that can be acquired by the morally correct exercising of freedom of choice. This remains true despite the fact that all three of these predispositions (not just personality – this point will become important in section two of this paper) are "predispositions *to the good* (they demand compliance with it)" (AA RGV 06:28).

Nonetheless, these characteristics are, according to Kant, necessary even for the *possibility* of human nature. Thus, presumably, these characteristics are contained within the concept of human nature on Kant's account and can be analyzed out of it. In any event, Kant does not offer further proof of this part of his moral psychology. Let us turn, then, to Kant's presentation of "the propensity to evil in human nature" (AA RGV 06:28), which is only *contingent* (AA RGV 06:29), and which is a moral judgment on the innate disposition of the human species. This is (as a moral judgment) what Kant claims in the introduction to R1 that he wants to prove.

[7] For discussions of the distinction between the predispositions to humanity and to personality, see Paul Guyer, *Kant on Freedom, Law and Happiness* (Cambridge University Press: 2000), 192n16; or Samuel Kahn, "Reconsidering 6:26n and the meaning of 'humanity'," in *Kant und die Philosophie in weltbürgerlicher Absicht* (Walter de Gruyter: 2013).

At this point in the text things get more complex. First, Kant adds in new terminology: "the will's capacity or incapacity arising from this natural propensity to adopt or not to adopt the moral law in its maxims can be called *the good or evil heart*" (AA RGV 06:29). This will become useful later. He next goes on to define the "three different grades of this natural propensity to evil" (AA RGV 06:29): frailty (AA RGV 06:29), impurity (AA RGV 06:30), and depravity (AA RGV 06:30). This is important because it shows that the stark binary above between good and evil is, on Kant's account, consistent with the idea that evil comes in degrees. Even more important for my purposes is that it is in this section of the text that Kant claims that "the propensity to evil is here established (as regards actions) in the human being, even the best" (AA RGV 06:30). This could be taken to mean either that the preceding paragraphs established that human beings have a propensity to evil (in the sense of a proof), or that this propensity to evil is established in these three propensities in any human being (in the sense of "found," or perhaps even "defined").

However, if it is to be the former interpretation, it is not entirely clear how the preceding definitions established this claim. Moreover, given the immediate context of this assertion, it seems as though the second interpretation is more reasonable.[8] That is, the sentence continues after a semicolon with a conditional: "*if* it is to be proved that the propensity to evil among human beings is universal…" (AA RGV 06:36). This indicates that Kant does not take himself to have proved that humans have a propensity to evil yet. Moreover, in the following paragraph, Kant distinguishes between juridical and

[8] Note, however, that this does not explain the footnote at AA RGV 06:39, which would seem to contradict this assertion: "the appropriate proof of this sentence of condemnation by reason sitting in moral judgment is contained not in this section, but in the previous one. This section contains only the corroboration of the judgment through experience." Although this could, indeed, be interpreted in such a way that it would be consistent with my assertion above (and the immediate context as well as the import of the rest of R1 (cf. e.g. the 'anthropological research' quotation from the remark at the beginning of R1, cited below) could support such an interpretation) I hesitate to say anything decisive here.

moral goodness – i.e. actions that merely correspond to the "letter of the law" as opposed to actions done from respect for the law (AA RGV 06:30-31).[9] This indicates that Kant is describing the ways in which evil is found in humans, not proving that there is some deep-seated attraction to it in our moral psychology. Finally, in the next section of the text, Kant continues as though the proof that human beings have a propensity to evil is yet to come – the proof of the judgment that human beings are evil by nature is yet to come in section three of the text.

Before presenting this proof, however, Kant elucidates what he means by his claim: "the statement, "the human being is *evil*," cannot mean anything else than that he is conscious of the moral law and yet has incorporated into his maxim the (occasional) deviation from it" (AA RGV 06:32). But this just means that the human being does not have a holy will,[10] and the contrast between the human being and a holy will is present even as early as the *Groundwork for a Metaphysics of Morals* and remains in the *Critique of Practical Reason* as well as in the *Lectures on Ethics* and the *Metaphysics of Morals*.[11] Indeed, this is what underlies the difference between a practical law and the way that we (as human beings) cognize this practical law: as an imperative. On Kant's account, the moral law in its pure form is cognized as such only by a divine will; we cognize the moral law as a constraining "ought" on our recalcitrant wills (AA MdS 06:222). Kant's assertion, then, that the human being has a propensity to evil seems to be nothing more than that we are not divine wills – and this is consistent throughout his writing. Moreover, it seems reasonable: it is unlikely that anyone would dispute the claim that we are not omnibenevolent.

[9] See also the second paragraph following the 'establishment' assertion: "the following elucidation is *also* necessary in order to *define* the *concept* of this propensity" (AA RGV 06:31, my emphasis).

[10] For an alternative reading, see Pablo Muchnik, "On the alleged vacuity of Kant's concept of evil," *Kant-Studien* 2006, 97 (4): 430-451.

[11] See, e.g., AA GMS 04:397, 413, 426, 434, 439, 449; AA KpV 05:25, 32, 82; AA Vorl 27:489; AA MdS 06:222, etc.

But reasonable is not grounds for a formal proof, and we must therefore wonder why Kant claims that "we can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being in view of the multitude of woeful examples that the experience of human *deeds* parades before us" (AA RGV 06:32-33). Perhaps he thought that his 'formal proof' (if, indeed, such a proof is even possible) might be misconstrued as proving that this evil is necessary in human beings in the sense that it cannot be avoided though it can still be imputed. The worry would be that in light of such a proof, this propensity to evil might come to be regarded as grounds for asserting the futility of attempting to follow the rigid moral law and, thus, be grounds for compromising if not totally abrogating (since the first might indeed constitute the second insofar as the second is possible – more on this in section two of this paper) the moral law.[12] But Kant does not rest his omission of this proof on moralistic grounds: he explicitly states that this formal proof is unnecessary. How does he justify this claim?

Kant's reasoning for the sufficiency of an empirical proof of his claim lies in the preceding section where he claims that

> … the propensity to evil is a deed in the first meaning (*peccatum originarium*),[13] and at the same time the formal ground of every deed contrary to law according to the second meaning…[14] that resists the law materially (AA RGV 06:31).

In other words, the propensity is imputable insofar as it is a "deed." It is a "decision" made outside of time, and it is also the ground for any "woeful…human *deeds*" (AA RGV 06:33).

[12] See, e.g., AA RGV 06:51-52 and AA GMS 04:405.

[13] When he says "first meaning," Kant is referring to something he said immediately earlier in the text: "a subjective determining ground of the power of choice *that precedes every deed*" (AA RGV 06:31).

[14] When he says "second meaning," Kant is referring to something he said immediately earlier in the text: "the use of freedom through which the supreme maxim… is adopted in the power of choice… by which… actions themselves… are performed according to maxims" (AA RGV 06:31).

Moreover, because each deed is to be considered a free action,[15] in seeking its rational ground we can find it only in a propensity to evil. That is, any empirical evidence of action contrary to law is instantiation for the claim that the agent does not have a holy will (but not vice versa).

Kant recognizes that this proof is only empirical and, therefore, somewhat unsatisfactory. We cannot ever properly deduce someone's (even our own) maxims from anything, so we cannot properly deduce evil maxims from empirical actions.[16] Indeed, at AA RGV 06:37 he himself claims that "*if* a propensity to this does lie in human nature" – a conditional, not an indicative – and he himself observes at AA RGV 06:38-39 that "everyone can decide by himself" whether humans are by nature evil. However, at this point we might realize that a formal proof is unnecessary for what Kant wants to do. This propensity is *contingent*, not necessary.[17] Thus, all that Kant needed to show was that 1. we have a predisposition to the good (what this means and why Kant needs to argue this will be addressed in section two) and 2. that given this predisposition to the good, his theory can accommodate both evil and good dispositions and, moreover, that both of these can be *imputable*.[18] As he states in the remark at the beginning of R1: "[whether or not humans are by nature good or evil will be demonstrated later] if it transpires from *anthropological research* that the

[15] "[E]very evil action must be so considered… as if the human being had just fallen into it directly from the state of innocence" (AA RGV 06:41).

[16] This kind of claim is familiar from the *Groundwork for a Metaphysics of Morals* (AA GMS 04:390, 406-407). But it also shows up in the *Critique of Practical Reason* (AA KpV 05:47, 81) and the *Religion within the Boundaries of Mere Reason* (AA RGV 06:21).

[17] The contingency of actual evil is discussed in some detail in Paul Formosa, "Kant on the radical evil in human nature," *Philosophical Forum* 2007, 38 (3): 221–245. McMullin claims that radical evil is necessary for the transition into agency (which Kant refers to as the transition toward *Mündigkeit* in his essay *What is Enlightenment*), arguing that *this* kind of necessity does not compromise its imputability (Irene McMullin, "Kant on Radical Evil and the Origin of Moral Responsibility," *Kantian Review* 2013, 18 (1): 49-72).

[18] For an alternative reading, see Seiriol Morgan, "The missing formal proof of humanity's radical evil in Kant's religion," *Philosophical Review* 2005, 114 (1): 63-114.

grounds that justify us in attributing one of these two characters to a human being as innate are of such a nature that there is no cause for exempting anyone from it, and that the character therefore applies to the species" (AA RGV 06:25-26, my emphasis). The point here seems to be that anthropological research will be needed to establish the reliability of the inference of bad maxims from the existence of bad actions.[19] Perhaps Kant's idea is that only empirical research will be able to establish whether humans are sufficiently similar to warrant the kind of "maxim guessing" that most of us do on an everyday basis without even thinking about it, even when our conclusions shed a negative light on the moral characters of our interlocutors.[20]

Having accomplished this goal insofar as he had set out to (even if it rests on an empirical proof, we were told that it would from the beginning of R1) only two short tasks remain for Kant in R1: (1) to explain how this evil originated in such a way that it can still be imputed to the agent, and (2) to explain what his conception of evil involves. I will briefly discuss both of these before transitioning into section two of this paper.

Because the origin of evil in human nature must be imputable on Kant's account, it cannot be determined by laws of nature. Rather, it must be "bound… according to laws of freedom" (AA RGV 06:39), which means that "it cannot be derived from some *preceding* state or other" (AA RGV 06:39-40). The idea of a *temporal* origin of free actions and, therefore, of the propensity to evil is thus contradictory (AA RGV 06:40): the "ground of the *exercise* of freedom… must be sought in the representations of reason alone" (AA RGV 06:40). However,

[19] This kind of inference (and its presence in the *Religion within the Boundaries of Mere Reason*) is discussed at length in Carl Manrique, "Radical Evil and the invisibility of moral worth in Kant's *de Religion*," *Ideas y valores* 2007, 56 (135): 3-27.

[20] Here I agree with Grimm, who thinks that Kant's case for radical evil is anthropological at heart (Stephen R. Grimm, "Kant's argument for radical evil," *European Journal of Philosophy* 2002, 10 (2): 160–177. For an alternative account, see, e.g., Stephen Palmquist, "Kant's quasi-transcendental argument for a necessary and universal evil propensity in human nature," *Southern Journal of Philosophy* 2008, 46 (2): 261-297).

since we have a predisposition to the good, there is "no conceivable ground for us… from which moral evil could first have come in us" (AA RGV 06:43). We are, indeed, left in the dark in regard to this, but we are still to hold a person who commits an evil deed "at the moment of action just as accountable… as if… he had just stepped out of the state of innocence into evil" (AA RGV 06:41). Since we "cannot inquire into the origin in time of this deed" (AA RGV 06:41) we must "inquire only into its origin in reason" (AA RGV 06:41). And just what is this origin – what is it that evil actually involves on Kant's view?

Kant tells us that "[w]hether the human being is good or evil, must not lie in the difference between the incentives that he incorporates into his maxim (not in the material of the maxim) but in their *subordination* (in the form of the maxim): *which of the two he makes the condition of the other*" (AA RGV 06:36). The two incentives here are the incentive of self-love and the incentive of morality (i.e. the moral law). Whether somebody is good or evil, then, depends not simply on whether they gratify the incentives of self-love or not. It depends on *the way* in which these incentives are gratified. If they are given superior weight than the moral law – if the moral law is fulfilled only conditionally based on the incentive of self-love – then the human being is evil. If, in contradistinction, the moral law is made the condition of fulfilling any incentive of self-love – if duty is always performed first and foremost – then the human being is good. In other words, "inclinations only make more difficult the *execution* of the good maxims opposing them; whereas genuine evil consists in our *will* not to resist the inclination when they invite transgression" (AA RGV 06:59n): the evil is not to be sought in inclinations and, therefore, in contrast to the stoics (as Kant portrays them), Kant would not have us eliminate them. Kant thinks that the correct response to the pull of the inclinations is to subordinate them to the moral law rather than to (try to) crush them into nonexistence.

It is this that underlies Kant's claim at AA RGV 06:58, at the opening of book II of the *Religion within the Boundaries of Mere Reason*, that "*considered in themselves* natural inclinations are *good*" and that to extirpate them would thus be both "harmful and blameworthy" (AA RGV 06:41). This seems to be in keeping with what Kant claimed earlier – that *all three* predispositions (not just personality) promote the good (AA RGV 06:28 (cited above)). However, this claim might be somewhat misleading: Kant is not claiming that the inclinations are unconditionally good, much less that their fulfillment is. He is not claiming that there is, in the inclinations, as in a good will, an incommensurate, absolute worth that is good in any and all situations. Rather, he is claiming that "other things being equal, the fulfillment of human inclinations can be assumed to be a part of what is good for human beings."[21] What is key is the "other things being equal" clause: we are not, of course, supposed to subordinate the moral law to this goal, but this does not mean that we cannot pursue it. In other words, the inclinations and the pursuit of happiness are not merely *instrumentally* good – but they are nonetheless not *unconditionally* good.[22]

Section 2: some problems in need of reconciliation

Based on the previous section, it can be seen that in the *Religion within the Boundaries of Mere Reason*, moral evil is a "*radical* innate *evil* in human nature" (AA RGV 06:32) because it is an evil of free choice: it is imputable. The grounds of this evil are thus not to be sought in a "sensuous nature" (AA RGV 06:34-35) because these "bear no direct relation to evil" (AA RGV 06:35). A human's inclinations are, in a sense, outside of the realm of good and evil because they are outside of freedom and, thus, outside of the moral law and

[21] Cf. Paul Guyer, *Kant on Freedom, Law, and Happiness* (Cambridge University Press: 2000), p. 224.
[22] Cf. Allen Wood, *Kant's Ethical Thought* (Cambridge University Press: 1999), section 9.4.

therefore the basis of any moral value judgment. Moreover, we "cannot presume ourselves responsible for their existence" (AA RGV 06:35); they are not imputable because they are outside of our sphere of action as free beings. We can, indeed, decide whether or not to subordinate our inclinations to the moral law, but we cannot decide (in any nontrivial fashion) not to have natural inclinations – and this is not a bad thing because the natural inclinations are not, in themselves, the ground of evil.

However, in the *Groundwork for a Metaphysics of Morals*, the picture seems to be somewhat different. As Rawls notes in *Kant IX*,[23] when Kant is speaking of the "hardened criminal" in section III of the *Groundwork for a Metaphysics of Morals*, he states that

> This better person, however, he believes himself to be when he transports himself into the standpoint of a member of the world of understanding, to which the idea of freedom, i.e., independence of *determining* causes of the sensible world, involuntarily necessitates him, and in which he is conscious of good will, which constitutes by his own admission the law for his evil will as a member of the sensible world, the law with whose authority he becomes acquainted when he transgresses it. The moral 'ought' is thus his own necessary volition as a member of an intelligible world and is thought of by him as an 'ought' only insofar as he at the same time considers himself as a member of the sensible world (AA GMS 04:454-455).

According to this passage, the moral, "I ought" becomes an indicative, "I will," in the world of the understanding. But this

[23] John Rawls, *Lectures on the History of Moral Philosophy* (Harvard University Press: 2000), p. 303. This issue is discussed by many other commentators, too. See, for example, Henry Allison, *Kant's Theory of Freedom* (Cambridge University Press: 1990); Lewis White Beck, *A Commentary on Kant's Critique of Practical Reason* (University of Chicago Press: 1960); Robert Louden, *Kant's Impure Ethics* (Oxford University Press: 2000); or H. J. Patton, *The Categorical Imperative* (University of Pennsylvania Press: 1971).

leaves a person *qua* intelligible being *no choice* but to be good. More problematic still, a person *qua* sensible being is entirely driven by causal impulses that are beyond his/her control. This is in line with what Kant says in section II of the *Groundwork for a Metaphysics of Morals*: "the inclinations themselves, however, as sources of needs, are so little of absolute worth… that rather to be entirely free of them must be the universal wish of every rational being" (AA GMS 04:428). Indeed, even in section I we find Kant pitting reason directly against the inclinations and happiness: the exercise of reason gives rise to "misology" (hatred of reason) because the more it is used "the more the human being falls short of true contentment" (AA GMS 04:395).

It is from these kinds of claims that what I have called the "Sidgwick problem" (in the introduction to this paper) comes.[24] One way to think about this is as follows. Since "everything in nature works in accordance with laws" (AA GMS 04:412), considering myself as a sensible being, no evil can be imputed to me. But in the intelligible world, the law of freedom is now a causal acausal law and so, considering myself as an intelligible being, no evil can be imputed to me. Therefore, considering myself either as a member of the world of the understanding or as a member of the world of nature I can be neither good nor evil since, even if I am following the moral law, my actions cannot be imputed to me.[25]

[24] Korsgaard raises a similar problem in chapter 7 of *Creating the Kingdom of Ends* (Christine Korsgaard, *Creating the Kingdom of Ends* (Cambridge University Press: 1996), chpt. 7). She points out that based on the account in the *Groundwork for a Metaphysics of Morals*, an all or nothing dilemma arises: considered as noumena, agents are responsible for all of their actions, whereas considered as phenomena, agents are not responsible for any of their actions. According to Korsgaard, this is counterintuitive. It is counterintuitive because sometimes we excuse agents for one or more actions on empirical grounds, and we do not think that this in any way compromises their agenthood. Although this problem is slightly different from what I am calling the "Sidgwick problem" above, hopefully it is clear that both problems derive from the same root.

[25] See also AA KpV 05:87, where Kant seems to make a similar claim in the *Critique of Practical Reason*.

I have called this the "Sidgwick problem" because Henry Sidgwick articulated it in *The Methods of Ethics*. It should be noted that Reinhold raised almost exactly the same problem a century earlier. However, when Reinhold pointed out this problem, it was not addressed primarily to Kant or Kant's philosophy *per se*. Rather, Reinhold thought the problem arose because of a *mis*interpretation of Kant that was facilitated by someone else: Schmid. To explain, Schmid had published a Kant lexicon in which he defined freedom as "dependency of the will on reason that immediately determines it."[26] Reinhold pointed out that this entails that immoral actions are not free and, thus, are not imputable, which is absurd. However, Reinhold argued that the absurdity lay in Schmid's definition rather than in Kant: Schmid had given an incorrect definition of Kant's concept of freedom. In other words, Reinhold argued that Schmid had misunderstood Kant's account.[27]

Sidgwick, by way of contrast, poses the problem directly to Kant. He argues that Kant makes a mistake because "the life of the saint must be just as much subject – in any particular portion of it – to the necessary laws of physical causation as the life of the scoundrel."[28] In other words, according to Sidgwick at least some things Kant says in the *Groundwork for a Metaphysics of Morals* (and the *Critique of Practical Reason*) seem to entail that Kant's ideas about freedom at this point require more working out. And this should be unsurprising: Kant's first mature stab at the problem of free will had come only a few years earlier – in 1781 in the third antinomy of the *Critique of Pure Reason*. At that point, Kant did not anticipate having to write anything like the *Groundwork for a Metaphysics of*

[26] *The following attributions were added post-publication on 26/04/2014:* This translation from M. Schmid, *Wörterbuch zum leichtern Gebrauch der Kantischen Schriften*, 2nd, expanded, edn. Jena. 1788: p.223, is by Courtney Fugate. See also Courtney Fugate's full discussion in his article, "On a supposed solution to the Reinhold/Sidgwick problem in Kant's *Metaphysics of Morals*", *The European Journal of Philosophy* (Virtual Issue: 2012).

[27] Karl Reinhold, *Letters on the Kantian Philosophy*, edited by Karl Ameriks and translated by James Hebbeler (Cambridge University Press: 1995).

[28] Henry Sidgwick, *The Methods of Ethics, 7th edition* (Hackett: 1981), p. 516.

Morals. Similarly, when Kant wrote the *Groundwork for a Metaphysics of Morals* in 1785, he did not anticipate having to write the *Critique of Practical Reason* – which he nonetheless published only three years later.

Kant's moral philosophy, like much of his philosophy, was not developed according to a pre-existing, fully worked out plan. As we have seen already, Kant corrects for the Sidgwick problem in R1 by making evil arise out of freedom of choice. Evil arises out of a misprioritization rather than out of our sensuous nature directly because if it arose out of our sensuous nature directly – out of our natural inclinations – then it would not be imputable. It is certainly true that the seeds for this theory can be found already in the *Groundwork for a Metaphysics of Morals*. For example, in the opening lines of section III of the *Groundwork for a Metaphysics of Morals*, Kant distinguishes between two different kinds of freedom, negative freedom and positive freedom.[29] Perhaps it was working this distinction out that led him to the theory he sets out in R1. But the point is that he does not work this distinction out – he does not use this distinction to good effect – at the time of the *Groundwork for a Metaphysics of Morals*.[30] The point is that in 1785, these ideas are mixed with other ideas, ideas about the causality of reason and of inclination, ideas that are ultimately dropped. Indeed, the theory of evil that Kant begins developing in the *Religion within the Boundaries of Mere Reason* does not seem to reach fruition until 1797, when Kant published the *Metaphysics of Morals*, in which he argues that some parts of our sensuous nature are just as necessary for morality as the faculty of reason.[31]

Before turning away from the Sidgwick problem, I want to indicate where I fit into the broader debate under which discussion of this problem usually falls: the *wille-willkür* debate. According to one group of commentators, the Sidgwick

[29] AA GMS 04:446-447.
[30] For a similar reading, see Matthew Caswell, "The value of humanity and Kant's conception of evil," *Journal of the History of Philosophy* (44.4, 2006): 635-663.
[31] See, e.g., AA MdS 06:399-403.

problem arises from the simple fact that Kant had not developed the correct terminology to talk about his position until the *Religion within the Boundaries of Mere Reason* and, even more so, the *Metaphysics of Morals.* These commentators argue that the account of immoral behavior (and the account of freedom it presupposes) that was laid out in the previous section of this paper is latent throughout Kant's corpus even though it was not made explicit until some of his later works. Thus, these commentators think that Schmid misinterpreted Kant, that Sidgwick did, too, and that Reinhold (and 1790s Kant) set things aright.[32]

According to a second group of commentators, this first group is incorrect. They argue that, in fact, Schmid had things right all along: freedom and morality *do* go hand in hand. Moreover, this is not so merely for the *Groundwork for a Metaphysics of Morals* and for the *Critique of Practical Reason*. This is so also for the *Religion within the Boundaries of Mere Reason* and for the *Metaphysics of Morals*. These commentators argue that evil is not merely unimputable; it is inconceivable. The solution to the puzzle is not located in redefining freedom. Instead, the solution to the puzzle is that Kant wants to redefine evil: evil is not acting according to inclination or elevating an immoral maxim to the status of a universal law. Rather, evil is making an exception of oneself to a law that one recognizes as authoritative. Moreover, these commentators argue that this is what Kant was saying all along. Reinhold messed things up because, by virtue of misunderstanding Kant's account of the nature of evil, he posed a pseudoproblem. Sidgwick did the same.[33]

I am dubious of both of these positions. As I noted above, I would agree with the first group of commentators that there are passages in the *Groundwork for a Metaphysics of Morals* and the *Critique of Practical Reason* that support the idea that

[32] See, for example, Nelson Potter, Jr., "Does Kant have two concepts of freedom?" *Akten des 4. Internationalen Kant-Kongress* (de Gruyter: 1974).
[33] Courtney Fugate, "On a supposed solution to the Reinhold/Sidgwick problem in Kant's *Metaphysics of Morals*," *The European Journal of Philosophy* (Virtual Issue: 2012).

Schmid and Sidgwick misinterpreted Kant.[34] However, there are also passages (like the one reproduced above) that suggest otherwise.[35]So I think that the first position is too extreme. But I think the second position is too extreme, too. It is certainly true that there are passages in the *Groundwork for a Metaphysics of Morals* and in the *Critique of Practical Reason* that suggest that Kant's early account of evil was more complex than the Sidgwick problem would require.[36]It also seems plausible that there are parts of the *Religion within the Boundaries of Mere Reason* and of the *Metaphysics of Morals* that suggest that a will is free if and only if it is moral.[37] However, this requires overlooking the passages in which evil is traced back to inclinations acting by causal laws. Moreover, I think the "solution" posed by the second position does not work: if the equation of freedom and morality is accepted, it does not matter how evil is defined. That is, provided that freedom is taken to be a

[34] One of the more important texts to which proponents of this interpretation point comes from the *Critique of the Power of Judgment*: "I deliberately say "under moral laws." The final end of creation is not the human being in accordance with moral laws, i.e., one who behaves in accordance with them" (AA KU 05:448n-449n). The idea is that in this passage Kant distinguishes between being under moral laws and acting in accordance with moral laws. This suggests that one can be under moral laws (i.e., be subject to moral laws) without acting in accordance with them. This is relevant because even in his earlier work Kant sometimes says that a free will is one that is under moral laws. This supports the idea that Kant never intended to say that a will is free if but only if it acts according to moral laws, evil if but only if it acts according to causal laws.

Fugate, who belongs to the second group of commentators, objects to the use of this passage in this debate. He argues that "[s]ince the very reason Kant gives for using 'under' instead of 'according to' in this passage is to put a limit on how much *nature* can be responsible for our moral status and not on how much *we* can, it provides no direct support for any claim about how freedom is related to the law" (Courtney Fugate, "On a supposed solution to the Reinhold/Sidgwick problem in Kant's *Metaphysics of Morals*," *The European Journal of Philosophy* (Virtual Issue: 2012), pp. 3-4). Although I do not agree with the first group of commentators (against whom Fugate is arguing), I think that Fugate misses the mark here. It really does not matter that in this footnote Kant is talking about how much nature can be responsible for with regard to our moral status. The point is simply that in this footnote, Kant distinguishes between being under moral laws and acting according to them, which suggests that it is possible to be under them while nonetheless not acting according to them.

[35] In addition to the passages cited above, see AA KpV 05:29, 33, 78, 96 or 98.

[36] See, e.g., AA GMS 04:424.

[37] See, e.g., AA MdS 06:382n.

precondition of imputability (which Fugate, the primary proponent of the second position, explicitly accepts)[38] the equation of freedom and morality entails that evil is not imputable regardless of its definition. So redefining evil is a nonstarter.

Thus, I align myself with a third group of commentators, whom I take to be more moderate.[39] Basically, the idea is that the texts are ambivalent. Sometimes, especially in his earlier writing (the *Groundwork for a Metaphysics of Morals* and the *Critique of Practical Reason*), Kant says that on the one hand, freedom and morality go together, and on the other hand, evil and being causally determined by inclinations go together. That is, sometimes the texts suggest that Kant steps directly into the trap of the Sidgwick problem. Other times, especially in his later writing (the *Religion within the Boundaries of Mere Reason* and the *Metaphysics of Morals*) Kant suggests that the inclinations are actually important for moral behavior just as much as for immoral behavior; thus, sometimes the texts suggest that Kant neatly sidesteps the Sidgwick problem, saying that negative freedom (the capacity for acting according to the moral law) and imputability go together and evil arises not directly from the inclinations but rather from a free choice that involves a misprioritization, an elevation of self-love over the moral law.

Closely related to the Sidgwick problem is what I have called the "Dignity and Worth" problem. In elucidating why good and evil must arise out of free choice in subordinating our incentives in our maxims, Kant also claims that the ground of evil cannot lie "in a *corruption* of the morally legislative reason, as if reason could extirpate within itself the dignity of the law itself" (AA RGV 06:35). In other words, our personality – "the susceptibility to respect for the moral law *as of itself sufficient incentive to the power of choice*" (AA RGV 06:27) – cannot have any

[38] Courtney Fugate, "On a supposed solution to the Reinhold/Sidgwick problem in Kant's *Metaphysics of Morals*," *The European Journal of Philosophy* (Virtual Issue: 2012), pp. 17*ff*.

[39] In this I take myself to be following more or less in the footsteps of Rawls, Wood and others.

evil grafted onto it. We cannot, as Wood observes,[40] as Lucifer in Milton's *Paradise Lost*, declare, "evil, be thou my good" (IV, 108) because this would mean that we were no longer subject to the moral law. We cannot be a "*diabolical* being" because this is simply inconceivable since it "would amount to a cause operating without any law at all… [which] is a contradiction" (AA RGV 06:35). And thus how are we to represent evil? Since evil arises from freedom of choice (thus not sensible nature or, as we have just seen, a corruption of our intelligible nature), evil should be named "*perversity* of the heart, and this heart is then called *evil* because of what results. An evil heart can coexist with a will which in the abstract is good" (AA RGV 06:37). And how does this fix some earlier problem in the *Groundwork for a Metaphysics of Morals*?

Kant opens section I of the *Groundwork of a Metaphysics of Morals* with his famous claim that "there is nothing it is possible to think of anywhere in the world, or indeed anything at all outside it, that can be held to be good without limitation excepting only a **good will**" (AA GMS 04:393). But the most natural way of interpreting this is as an emphatic claim that our worth – our dignity – is, in fact, contingent on our being morally good. That is, if a good will is good without limitation, then a person with a good will has something good without limitation. Thus, it seems, a morally good person has absolute worth and dignity. But if nothing other than a good will is good without limitation, then a person who does not have a good will has nothing that is good without limitation. Thus, it seems, an evil person does not have absolute worth or dignity.

In section II, Kant seems to say otherwise when he posits his formula of humanity, which is grounded on the absolute worth of a human being *qua* person. But he goes on to claim that we all have an incommensurable dignity that arises from "the idea… of a rational being that obeys no law except that which at the same time it gives itself" (AA GMS 04:434). Given that

[40] Allen Wood, *Kant's Ethical Thought* (Cambridge University Press: 1999), p.373n3.

in the *Groundwork for a Metaphysics of Morals* Kant seems to think that an agent gives a law to him/herself only if s/he is acting morally (AA GMS 04:440-442), the most natural way of interpreting this is, again, as a claim that our dignity is founded on our rationality – actual rationality, not capacity for rationality. Indeed, even as late as the *Critique of the Power of Judgment* Kant seems to advance ideas that are consistent with this: "only through that which he does without regard to enjoyment, in full freedom and independently of that which nature could passively provide for him, does he give his being as the existence of a person an absolute value" (AA KU 05:208-209).

There are ways in which one could try to massage these value claims into consistency. For example, Korsgaard argues that the value of the capacity for rationality is linked to the value of rationality because the capacity for rationality, "completed and perfected," becomes rationality.[41] This link seems undeniably true on Kant's account. Nonetheless, as a "rescue attempt," this seems doomed to failure: there seems very little doubt that in the passages noted above, if the bare capacity for rationality has value, this value is *derivative* of actual rationality. Perhaps this "massaging" of the texts also requires some looking away. One thing that this reveals is that Korsgaard's famous regress argument, which makes the value of the capacity for rationality logically prior to the value of rationality, is, at least from an exegetical perspective, oversimplified.[42] Indeed, in recent work commentators like Kerstein have taken Korsgaard to task on exactly this front, arguing that a better interpretation of Kant's argument for the formula of humanity at AA GMS 04:428-429 yields the duty

[41] Christine Korsgaard, *Creating the Kingdom of Ends* (Cambridge University Press: 1996), p. 114.

[42] I argue that the regress argument requires eschewing the distinction between culpable and nonculpable ignorance in my Samuel Kahn, "The Guise of the Objectively Good," *The Journal of Value Inquiry* (47.1-2 (June 2013)): 87-99.

always to treat rationality (rather than the capacity for rationality) as an end and never merely as a means.[43]

However, my goal here is not to get involved in disputes about Kant's argument for the formula of humanity. Rather, I want simply to point out that in R1 Kant has gotten around once and for all the problem raised by the famous first line of the *Groundwork for a Metaphysics of Morals*: in R1, he claims that we *always* have a good will (albeit only in the abstract). What is at question is, rather, whether we have a good *heart*. Moreover, since we always have a good will, our dignity (and, thus, our place in the formula of humanity as absolute ends) always remains intact regardless of our virtue or viciousness: our absolute worth is not, in point of fact, dependent on whether or not we are morally good. Rather, we always have worth through our possession of a good will in the abstract, even when we have an evil heart. And then what is this "good will in the abstract," this worth of personality, in fact based on? It seems it can be nothing other than the capacity for autonomy and positive freedom (AA RGV 06:223).

Section 3: conclusion

To recapitulate, in R1 Kant offers a conception of freedom of choice whereby we have a predisposition to good and a propensity to evil. In so doing, he claims that moral good or evil arises not out of either our sensible or our intelligible nature *per se* – rather, it arises out of our freedom of choice directly: we are good if we subordinate our natural inclinations to the moral law and evil if we subordinate the moral law to our natural inclinations. He claimed to show from anthropological research that we have a propensity to evil. However, this propensity is only a *contingent* disposition in us; it still arises from freedom of choice and is still imputable.

[43] Samuel Kerstein, "Deriving the Formula of Humanity (GMS II, 427-437)," in *Groundwork for the Metaphysics of Morals* (Walter de Gruyter: 2006).

Moreover, what is at question is not whether or not we have a good will – it is whether or not we have a good heart.

By avoiding making sensible nature in itself the ground of evil and intelligible nature the ground of good (as he seems to do in the *Groundwork for a Metaphysics of Morals*), Kant avoids two closely related but nonetheless distinct problems: the Sidgwick problem and what I have called the Dignity and Worth problem. If we are evil it does not come from our being determined by sensibility – it comes from our subordination of the moral law to sensibility. Thus evil (and goodness) is imputable. Moreover, even if we are evil this does not compromise our absolute worth as human beings: our dignity is incommensurable and we preserve our status as absolute ends because we always have a good will (even if only in the abstract) because this is based on the ideas of freedom and autonomy – things that we simply cannot (in any nontrivial fashion) forego.

This being said, we might still object to Kant's view on another ground. We might wonder, if human beings do have a contingent propensity to evil – if Kant's empirical claims are accurate – and if, as Kant claims in the introduction to R1, we can claim empirically that the "history of the world attests too powerfully against [the view that the human species is improving morally]" (AA RGV 06:19), what does this mean for Kant's arguments for the practical postulates as posited in, for example, the *Critique of Practical Reason*?[44] Let me explain.

In the *Critique of Practical Reason*, Kant argues that we have a moral duty to promote a world in which everyone is maximally virtuous and in which happiness is distributed in accordance with virtue. Because we ought to do this, we must be able to do so. But in order to do so rationally, we must take such a world to be not merely logically possible (i.e. free of internal contradiction) but also really possible (i.e. to have a

[44] The beginnings of this connection are discussed in Matthew Caswell, "Kant's conception of the highest good, the Gesinnung, and the Theory of Radical Evil," *Kant-Studien* 2006, 97 (2): 184-209.

ground in reality). For example, I take perpetual motion machines to be logically possible. However, it would be irrational for me to try to build one because I do not take them to be really possible (quite the contrary: I take them to violate plausible physical laws, like the 2nd law of thermodynamics). This is relevant here because in the *Critique of Practical Reason*, Kant claims that we must assume that we are immortal in order to make rational the pursuit of perfect virtue (which, because of our sensuous nature, because of our inclinations, can be attained only in an infinite amount of time) and that there is a God in order to make rational the pursuit of a world in which happiness is distributed in accordance with virtue (AA KpV 05:120).[45] In particular, the pursuit of such a world requires belief in the existence of a God who is omnipotent, omniscient and omnibenevolent in order to ensure that this God is aware of as well as able and willing to correct all mistakes in the distribution.[46]

The discussion in this paper raises two questions about this argument, one for each of the practical postulates. These questions cannot be discussed thoroughly here. But in the last paragraphs of this paper, I would like to gesture toward them.

The question for the postulate of immortality is whether it is necessary. That is, given Kant's change in the *Religion within the Boundaries of Mere Reason* – given that he no longer views the pursuit of morality as a striving against and perhaps even as a seeking to extirpate the inclinations – given that he now views the pursuit of morality as a seeking merely to subordinate the inclinations to the moral law – it is entirely unclear why perfect virtue could not be attained in the normal course of a human life. This is not to say that it ever is attained in the normal course of a human life. It is simply to say that it is unclear why

[45] The argument for the practical postulates in the "Canon of Pure Reason" chapter of the *Critique of Pure Reason* is largely the same although it differs in the details. In particular, the reasons why immortality and God must be assumed to make the pursuit of the highest good rational are slightly different. But I cannot pursue these details here.

[46] For a thorough discussion of Kant's practical postulates, see Allen Wood, *Kant's Moral Religion* (Cornell University Press: 1968).

it should not be, which seems to render the postulate of immortality moot. Indeed, this might explain the conspicuous absence of any discussion of immortality in Kant's discussion of the practical postulates in the *Critique of the Power of Judgment*, published only shortly before Kant wrote the *Religion within the Boundaries of Mere Reason* – or in the doctrine of method of the *Metaphysics of Morals* (which does hint at the postulate of the existence of God).[47]

The question for the postulate of the existence of God is whether it is sufficient. That is, given that a moral disposition is now within our reach and yet, if appearances be trusted, so rarely achieved, we might come to believe that even if there is a supernatural deity, this deity is not much concerned with our petty grievances or with the advancement of our all too human conception of the highest good. This might explain why, beginning with the *Critique of the Power of Judgment*, Kant seems to edge away from the postulate of the existence of God, an "edging" that is carried still further in the *Metaphysics of Morals* and the *Opus Postumum*. However, I cannot discuss these questions in detail here.

Bibliography

Allison, Henry, 'On the very idea of a propensity to evil', *The Journal of Value Inquiry* 2002, 36: 337-348.

Allison, Henry, *Kant's Theory of Freedom* (Cambridge University Press: 1990).

Beck, Lewis White, *A Commentary on Kant's Critique of Practical Reason* (University of Chicago Press: 1960).

Caswell, Matthew, 'The value of humanity and Kant's conception of evil', *Journal of the History of Philosophy* (44.4, 2006): 635-663.

Caswell, Matthew, 'Kant's conception of the highest good, the Gesinnung, and the Theory of Radical Evil', *Kant-Studien* 2006, 97 (2): 184-209.

[47] Cf. Allen Wood, *Kant's Moral Religion* (Cornell University Press: 1968).

Formosa, Paul, 'Kant on the radical evil in human nature,' *Philosophical Forum* 2007, 38 (3): 221–245.
Fugate, Courtney, 'On a supposed solution to the Reinhold/Sidgwick problem in Kant's *Metaphysics of Morals*', *The European Journal of Philosophy* (Virtual Issue: 2012).
Grimm, Stephen R., 'Kant's argument for radical evil', *European Journal of Philosophy* 2002, 10 (2): 160–177.
Guyer, Paul, *Kant on Freedom, Law and Happiness* (Cambridge University Press: 2000).
Kahn, Samuel, 'Reconsidering 6:26n and the meaning of "humanity"', in *Kant und die Philosophie in weltbürgerlicher Absicht* (Walter de Gruyter: 2013).
Kahn, Samuel, 'The Guise of the Objectively Good', *The Journal of Value Inquiry* (47.1-2 (June 2013)): 87-99.
Kant, Immanuel, *Kants gesammelte Schriften*, 29 vols., issued by the Prussischen Akademie der Wissenschaften (vols. 1–22), the deutschen Akademie der Wissenschaften (vol. 23), and the Akademie der Wissenschaften zu Göttingen (vols. 24–9). (de Gruyter: 1902-).
Kant, Immanuel, *Lectures on Ethics*, ed. by Peter Heath and J.B. Schneewind and trans. by Peter Heath (Cambridge University Press: 1997).
Kant, Immanuel, *Critique of Pure Reason*, trans. and ed. by Paul Guyer and Allen Wood (Cambridge University Press: 1998)
Kant, Immanuel, *Practical Philosophy*, trans. and ed. by Mary J. Gregor and intro. by Allen Wood (Cambridge University Press: 1996).
Kant, Immanuel, *Religion and Rational Theology*, trans. and ed. by Allen Wood and George di Giovanni (Cambridge University Press: 1996).
Kant, Immanuel, *Critique of the Power of Judgment*, ed. by Paul Guyer and trans. by Paul Guyer and Eric Matthews (Cambridge University Press: 2000).
Kerstein, Samuel, 'Deriving the Formula of Humanity (GMS II, 427-437)', in *Groundwork for the Metaphysics of Morals* (Walter de Gruyter: 2006).
Korsgaard, Christine, *Creating the Kingdom of Ends* (Cambridge University Press: 1996).
Louden, Robert, *Kant's Impure Ethics* (Oxford University Press: 2000).
Manrique, Carl, 'Radical Evil and the invisibility of moral worth in Kant's *de Religion*', *Ideas y valores* 2007, 56 (135): 3-27.
McMullin, Irene, 'Kant on Radical Evil and the Origin of Moral Responsibility', *Kantian Review* 2013, 18 (1): 49-72.
Morgan, Seiriol, 'The missing formal proof of humanity's radical evil in Kant's religion', *Philosophical Review* 2005, 114 (1): 63-114.

Muchnik, Pablo, 'On the alleged vacuity of Kant's concept of evil', *Kant-Studien* 2006, 97 (4): 430-451.
Palmquist, Stephen, 'Kant's quasi-transcendental argument for a necessary and universal evil propensity in human nature', *Southern Journal of Philosophy* 2008, 46 (2): 261-297.
Patton, H. J., *The Categorical Imperative* (University of Pennsylvania Press: 1971).
Potter, Jr., Nelson, 'Does Kant have two concepts of freedom?' *Akten des 4. Internationalen Kant-Kongress* (de Gruyter: 1974).
Rawls, John, *Lectures on the History of Moral Philosophy* (Harvard University Press: 2000).
Reinhold, Karl, *Letters on the Kantian Philosophy*, edited by Karl Ameriks and translated by James Hebbeler (Cambridge University Press: 1995).
Schmid, M., *Wörterbuch zum leichtern Gebrauch der Kantischen Schriften, dritte vermehrte Aufgabe* (Jena: 1795).
Sidgwick, Henry, *The Methods of Ethics, 7th edition* (Hackett: 1981).
Wood, Allen, *Kant's Moral Religion* (Cornell University Press: 1968).
Wood, Allen, *Kant's Ethical Thought* (Cambridge University Press: 1999).