Kant and the Problem of Moral Conversion

Ryan Kemp
Wheaton College

Abstract:

In this essay, I explore a series of issues related to Kant's account of moral conversion as it is developed in his late text *Religion within the Boundaries of Mere Reason*. In his attempt to explain how radically evil agents become good, Kant considers three conversion models that appear to mutually exhaust the explanatory options. I argue that Kant rejects them all and—precisely because the three models are exhaustive—concludes that the source of moral conversion is a mystery.

Introduction

Kant's *Religion within the Boundaries of Mere Reason* is a mixed bag. On the one hand, the development of the executive will (*Willkür*) and a dispositional account of moral evaluation (a moral *Gesinnung*) provides a welcome improvement over the flimsy moral-psychological framework offered in earlier texts like the *Groundwork*.[1] On the other hand, Kant makes a host of baffling philosophical claims, for instance, that a person's fundamental moral disposition is chosen "outside of time" and that this "choice" is *always* exercised in favor of evil (6:20, 22).[2] The latter—the idea that humans suffer from a kind of original sin—was particularly unpopular with Kant's contemporaries, leading Schiller to claim (a bit dramatically) that Kant's text had driven his "feelings into revolt."[3] Two centuries later versions of this basic complaint live on. Gordon Michalson has, for instance, recommended that Kant's account of radical evil should "be viewed as the final result of Kant's latent resentment against the body, his philosophical chagrin that pure reason must cohabit with sensuousness."[4] Though Michalson quite fortunately avoids Schiller's indignation, his assessment is no less dismissive.

In contrast to this more general tone of reproof, recent work on the *Religion* has made strides toward repairing the text's reputation. In this regard, two interpretive strategies stand out. The first, what we might call the "textual coherence" camp, has argued that aspects of the text typically considered problematic are, in fact, easily explained. Allen Wood, for instance, has attempted to naturalize Kant's account of radical evil by arguing that the latter is merely a social phenomenon—mere "unsocial sociability."[5] Other commentators, like Henry Allison, have worked to clarify the *Religion*'s

[1] Of course, some scholars think that the *Groundwork*'s account of agency already includes these elements. See, for instance, Allison 2011, 78-80; 296-300.

[2] All references to Kant are to the volume and page (e.g. 6:22) of *Kant's gesammelte Schriften*, herausgegeben von der Deutschen Akademie der Wissenschaften, Berlin: Walter de Gruyter, 1902ff. Translations are taken from *The Cambridge Edition of the Works of Immanuel Kant*, Cambridge: Cambridge University Press, 1998.

[3] Schiller 1992, 26:219.

[4] Michalson 1990, 69.

[5] Wood 1991. See also Anderson-Gold 1991.

argumentative structure by, for example, reconstructing Kant's promised (and skillfully hidden) "deduction" of radical evil.[6] Still others, focusing on Kant's claim concerning the universality of evil, have argued that radical evil is merely a regulative concept, one that is *regarded as* universal for the sake of moral striving.[7]

A second—no less friendly—interpretive approach focuses on the way in which claims put forward in the *Religion* play an integral role in the later work of the German Idealists. This second group of interpreters, what we might call the "textual inspiration" camp, is interested in highlighting the philosophical originality of the *Religion*. We see a recent example of this strategy played out in Samuel Loncar's essay "Converting the Kantian Self."[8] At the end of his paper Loncar suggests that perhaps the most important legacy of the *Religion* is not its status as a flawed text, but rather its status as a text that bequeaths a certain conception of agency: an account of autonomy that requires a "kind of *Selbst-Setzen,*" an original and spontaneous act of self-creation.[9] While, in this regard, Loncar only offers a promissory note, his suggestion opens up a potentially productive means of approaching the *Religion*.

In this essay, I begin the work of pursuing one such Kantian debt. Though, like Loncar, I take the *Religion's* account of moral conversion to be a productive starting point, I am not primarily interested in Kant's theory of agency. Instead, I plan to explore the way in which the problem of conversion itself is taken up by Kant's philosophical successors. On my view, what makes the *Religion*'s role in this story particularly important is that the later idealists adopt options that Kant himself rehearses and rejects. In this respect, the *Religion*'s account of conversion foreshadows fifty years of debate, one that takes us from Schelling and Fichte's early concerns about the transition from "dogmatism" to "idealism," through Hegel's account of "mediation" to Kierkegaard's picture of Christian "grace." In this essay, I lay the foundation for this larger narrative by outlining the three conversion models that Kant explores in the early pages of the *Religion*. Once we clarify what the models consist in and appreciate why Kant thinks they are flawed, we can begin to see how Kant's "scandalous" text both sets the stage for, and anticipates the outcome of, a larger idealist debate.

Section 1: The Problem of Moral Conversion (Some Stage Setting)

There is a popular reading of Kant's *Groundwork* that goes something like this: In ideal circumstances an agent's rational will exercises control over his actions; these reason governed actions are paradigmatically *autonomous*. In non-ideal circumstances, however, sensibility wrests control from the rational will and imposes its own external force; these inclination determined acts are paradigmatically *heteronomous*. Autonomous actions are morally praiseworthy and heteronomous acts are morally blameworthy.[10] Construed as such, it is not difficult to understand why Kant's contemporaries were unsatisfied with the *Groundwork*. Its account of the will appears to preclude

[6] Allison 2002. See also Palmquist 2008.

[7] See, for instance, Kemp 2011 and Muchnik 2010.

[8] Loncar 2013.

[9] Loncar 2013, 366.

[10] For a much subtler version of the same basic story, see Kosch 2006, 15-65.

immoral action: when an agent "transgresses" the moral law (i.e., when sensibility gains the upper hand) his will is *inactive;* he is not acting as much as he is being acted upon.[11]

Over the last thirty years, no scholar has fought harder to dispel the above story than Henry Allison. Allison has consistently and persuasively argued that the conceptual apparatus Kant develops in the *Religion* (namely, the distinction between the legislative and executive wills: *Wille/Willkür*) is present from the beginning of the critical project. Kant's introduction of the executive will is significant, because—unlike the legislative will which merely sets down the law—*Willkür* is responsible for actively adopting a particular course of action: choosing either to obey the dictates of *Wille* or reject them.[12]

Regardless of whether Allison's claims about the *Groundwork* are correct, Kant's *Religion within the Bounds of Mere Reason* quite clearly make a distinction between the will's legislative and executive capacities. There, Kant writes, "freedom of the power of choice [*Willkür*] has the characteristic, entirely peculiar to it, that it cannot be determined to action through any incentive [*Triebfeder*] *except so far as the human being has incorporated it into his maxim*" (6:23-24). Kant's claim that incentives are "incorporated" suggests that action involves more than merely affirming or rejecting a practical possibility. Incorporation involves endorsing a particular act insofar as that act falls under a more general principle of action. Thus, the act of, say, cheating a child out of his pocket money is *incorporated* into the more general principle of subverting the moral law when it conflicts with inclination. In the *Religion*, Kant refers to an agent's general principle of action as her "supreme maxim" [*oberste Maxime*] (6:31), that is, the maxim that underlies, and is reflected in, the particular choices of the agent.

For many commentators, this emphasis on an agent's supreme maxim signals a welcome shift in Kant's moral psychology. While in the *Groundwork* there is debate concerning whether the proper locus of moral evaluation is an agent's occurrent intentions or her underlying character (her *Denkungsart*),[13] the *Religion* makes it clear that the latter is primary: actions are morally significant only insofar as they reflect an agent's moral disposition (that is, her *Gesinnung*). As Allison notes, this shift places Kant in "partial agreement with a tradition in moral psychology that stretches at least back to Aristotle and that includes, in addition to Leibniz and Hume, contemporary thinkers who insist that moral responsibility presupposes that actions be connected with the character of an agent."[14] However, as Allison also notes, Kant distinguishes himself from the larger tradition by insisting that an agent freely chooses his underlying *Gesinnung*. Taking pains to distinguish between two different senses of the concept "deed," Kant explains: "Now, the term 'deed' [*Tat*] can in general apply just as well to the use of freedom through which the supreme maxim (either in favor of, or against, the law) is adopted in the power of choice, as to the use by which the actions [*Handlungen*] themselves

[11] For instance, Kant's contemporary Karl Reinhold writes in reference to the *Groundwork* account: "If the moral law announces to us no other freedom than that which consists in the self-activity of reason, then the ability to act *immorally* is not only an *inability*, but is simply *impossible*." This excerpt from Reinhold's *Auswahl vermischter Schriften* is reproduced in Bittner, R. and Cramer, K., *Materialien zu Kants Kritik der praktischen Vernunft*, Frankfurt/Main, Suhrkamp, 1975: 323-4.

[12] See especially Allison 1990, 85-145.

[13] See Ameriks 1989.

[14] Allison 1990, 137.

(materially considered, i.e. as regards the objects of the power of choice) are performed in accordance with that maxim" (6:32).

Kant's motivation for claiming that one's fundamental moral disposition is a product of a free choice [*Willkür*] is twofold. First, Kant thinks that predicating "good" or "evil" of an agent is tantamount to claiming that the agent is responsible for his character. Responsibility for one's character, however, assumes that one's character has been freely chosen—or so thinks Kant.[15] Second, Kant wants to distinguish his account of an original and "natural" moral state from the Christian account of original sin. Kant writes:

> But lest anyone be immediately scandalized by the expression *nature*, which would stand in direct contradiction to the predicates *morally* good or *morally* evil if taken to mean (as it usually does) the opposite of the ground of actions [arising] from *freedom*, let it be noted that by 'the nature of a human being' we only understand here the subjective ground…of the exercise of the human being's freedom in general (under objective moral laws) antecedent to every deed that falls within the scope of the senses" (6:21).

Kant's claim that a person's supreme maxim is "either in favor of, or against, the law" falls out of his account of human nature. Kant identifies three "predispositions to good in human nature" that are present in every human will. Humans possess (1) a predisposition to "animality," or an inclination to meet their physical needs; (2) a predisposition to "humanity," or an inclination to "gain worth in the opinion of others"; and (3) a predisposition to "personality," or a "susceptibility to respect for the moral law" (6:27). Though the predispositions are in themselves good and cannot be eradicated, the first two (the predispositions to animality and humanity) can be abused. When the predisposition to personality is subverted in favor of either of the others, a person is said to choose evil. Conversely, when the predisposition to personality rules over the other two, a person is said to choose the good. Since the particular arrangement of a person's moral *Gesinnung* is a product of a free choice, and the only available options are to give preference to moral personality or subvert it, Kant thinks that a morally indifferent supreme maxim is impossible. Insofar as a person has an *oberste Maxime*, they have taken a moral stand.

As I suggested above, one of the *Religion*'s more baffling claims comes in Kant's further suggestion that human beings originally (and always) choose to subvert the good. Kant infamously writes:

> "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…we can call this ground a natural propensity to evil, and, since it must nevertheless always come about through one's fault, we can further even call it a *radical* innate *evil* in human nature (not any the less brought upon us by ourselves…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 (6:32).

[15] He writes, "This disposition [i.e., one's *Gesinnung*] too…must be adopted through the free power of choice, for otherwise it could not be imputed" (6:25).

While many commentators have taken this and similar passages to suggest that Kant thinks all human beings necessarily choose evil,[16] I have recently argued against this view.[17] Kant is not, I suggest, making a claim about what all human beings *in fact* do; he is making a claim about how we *must view* human nature in light of our further commitment to moral striving. Moral striving requires that we presuppose an original state of evil. In this regard, attributions of universal evil are merely regulative. For the purposes of our current discussion, the importance of the "original evil" thesis lies in its connection to Kant's larger account of moral conversion. Kant spends much of the *Religion* thinking through how a person who starts off *radically* evil (that is, a person with a corrupted supreme maxim), can become a person who is radically good. The first hint that there may be an explanatory problem comes at the end of Part I, in Kant's "General Remark." He writes, "How is it possible that a naturally evil human being should make himself into a good human being surpasses every concept of ours. For how can an evil tree bear good fruit?" In a footnote, he poses the same problem from the other direction: "The tree, good in predisposition, is not yet good in deed; for, if it were so, it surely could not bring forth bad fruit" (6:45). While the uniquely "botanical" worry of producing good fruit from bad trees seems legitimately vexing (at least from where my armchair rests), we might wonder how well the tree analogy describes the problem of moral conversion in the context of Kant's moral psychology. Are humans like trees in this respect?

One source of disanalogy may lie in the fact that bad human beings (unlike bad trees) possess a predisposition to personality that is never eradicated. Kant even hints that this predisposition plays some role in the possibility of moral conversion. Just after the passage where Kant raises the problem of conversion, he writes: "Surely we must presuppose in all this that there is still a germ of goodness left in its entire purity, a germ that cannot be extirpated or corrupted" (6:45). Practically speaking, this suggests that an agent may be able to draw from uncorrupted elements of his nature in order to restore the diseased elements.

Kant goes on, however, to reject this possibility. Just as a healthy tree blossom cannot restore decaying roots, Kant indicates that the germ of goodness is not a *sufficient* resource for self-conversion. He writes, "This evil is *radical*, since it corrupts the ground of all maxims; as natural propensity, it is also not to be *extirpated* through human forces, for this could only happen through good maxims—something that cannot take place if the subjective supreme ground of all maxims is presupposed to be corrupted" (6:37). Why Kant thinks that supreme maxims delimit an agent's choices so absolutely is, I think, made more comprehensible when we compare his moral-psychological framework to a similar, more contemporary account: Harry Frankfurt's hierarchal account of the will.

For Frankfurt, an action is free if and only if it is caused by a lower order desire that conforms with an agent's highest order volition (that is, a desire for some lower order desire to be motivationally effectual). For Frankfurt, the most important highest order volitions are an agent's "loves" and "cares." An agent is never more autonomous (or free) then when he acts from his loves; he is never more heteronomous (or unfree) then when he acts to undermine them.[18] Here, then, we see a version of the problem that Kant is engaging with: if an agent deeply values, say, a life of selfish

[16] See Michalson 1990.

[17] Citation information redacted.

[18] Frankfurt 1998 and 1999.

hedonism, it "surpasses every concept of ours" how that agent can freely act to uproot his selfishness. Such a radically transformative act necessarily involves acting on a desire that conflicts with an agent's highest order value. Kant invites us to view the relationship between an agent's supreme maxim and the other aspects of his desiderative nature in a way that is similar to Frankfurt's portrayal of the relationship between an agent's loves and his first order desires. First order desires (like 'tree blossoms') do not have the practical authority to uproot the values that form the core of our practical identity ("tree roots").

Interestingly, this indicates that the "no-bad-action" problem that was earlier cited in reference to the *Groundwork*'s account of the rational will, is reintroduced in a new form in the *Religion*. Just as the agent of the *Groundwork* seems incapable of actively resisting the legislative will, the agent of the *Religion* cannot perform an action that undermines his supreme maxim: bad agents perform bad actions, while good agents perform good actions (a "no-*uncharacteristic*-action" problem). Kant addresses this issue in the context of explaining how to interpret acts that conform to the moral law at the empirical level. It is important that Kant address such putatively good acts insofar as their performance *seems* to show that good fruit can indeed come from bad trees. Kant writes: "In this reversal of incentives through a human being's maxim contrary to the moral order, actions [*Handlungen*] can still turn out to be as much in conformity to the law as if they had originated from true principles…The empirical character is then good but the intelligible character still evil" (6:36-37).

This distinction between the intelligible and the empirical character of an action is one that Kant first makes in the passage we read above where he differentiates the "deed" [*Tat*] of adopting a supreme maxim and the "deed" in which actions [*Handlungen*] are performed in accordance with the supreme maxim. Speaking of the relationship between the evil *Tat* and subsequent *Handlungen*, Kant writes: "the first indebtedness [that is, the *peccatum originarium*] remains even though the second [the *peccatum derivativum*] may be repeatedly avoided…The former is an intelligible deed, cognizable through reason alone apart from any temporal condition; the latter is sensible, empirical, given in time (*factum phenomenon*)" (6:31). This passage, like the one of the previous paragraph, suggests that the true bearer of moral value is always the original *Tat*. Even if an agent's *Handlungen* avoid evil, the original indebtedness is primary. As phenomenal acts, Kant claims that *Handlungen* can only attain to "legal" goodness; *moral* goodness pertains to an agent's supreme *Tat*. Kant continues, "However, that a human being should become not merely *legally* good, but *morally* good (pleasing to God) i.e. virtuous according to the intelligible character [of virtue] (*virtus noumenon*) and thus in need of no other incentive to recognize a duty except the representation of duty itself—that…cannot be effected through gradual *reform* but must rather be effected through a *revolution* in the disposition of the human being (a transition to the maxim of holiness of disposition)" (6:47).

Having now sketched the account of agency that undergirds Kant's account of conversion, we can see why the latter is so explanatorily problematic. The very problem that a moral *Gesinnung* is called upon to solve—namely, an account of how particular good or evil actions can be principled—creates a problem for explaining how an agent can reform himself. If acting from one's *Gesinnung* is a necessary condition for principled action, then action that is contrary to one's *Gesinnung* (for instance, uprooting one's current supreme maxim and replacing it with an opposed one) appears to be something that one cannot, in principle, do. This puzzle leads Kant to raise the question that frames our discussion: "[H]ow can an evil tree bear good fruit?" (6:45). Kant, however, continues:

> But, since by our previous admission a tree which was (in its predisposition) originally good but did bring forth bad fruits, and since the fall from good into evil (if we seriously consider that evil originates from freedom) is no more comprehensible than the ascent from evil back to the good, then the possibility of this last cannot be disputed. For, in spite of that fall, the command that we *ought* to become better human beings still resounds unabated in our souls; consequently, we must also be capable of it, even if what we can do is of itself insufficient and, by virtue of it, we only make ourselves receptive to a higher assistance inscrutable to us (6:45).

This passage is important because it makes clear that even though the mechanics of self-conversion are, in a sense, incomprehensible, morality demands that self-conversion is nonetheless possible: *ought implies can*. So when Kant, both in this and in subsequent passages, hints that "higher assistance" is the solution to the above problem, he cannot literally mean that God is the source of moral conversion. What, then, does he mean?

Section 2: Kant's "Solution" to the Problem of Moral Conversion

In this section of the essay, I explore three conversion models that Kant considers in Part I of the *Religion*: conversion by "spontaneous choice," conversion by "gradual reformation," and conversion by "divine assistance." My claim is not that Kant confusedly embraces each of these accounts, but rather that he entertains them as possibilities that he eventually rejects. The view that Kant settles on, what I call a "mystery" account of conversion, is not—philosophically speaking—any more satisfying than those he rejects. Kant's rejection of the three dominant theories, combined with his conviction that moral conversion must occur, leads him to a kind of agnosticism about the mechanics of conversion.

Section 2.1: Conversion by "Spontaneous Choice"

Recall from above that one of the major innovations of the *Religion* is Kant's explicit reference to the executive will or *Willkür*. The executive will is a powerful explanatory tool because it, unlike the legislative will, is capable of either rejecting or affirming the moral law. In Part I of the *Religion*, the executive will is put to good use in explaining how an agent originally chooses an evil *Gesinnung*. In an important passage Kant writes:

> This disposition…must be adopted through the free power of choice [*Willkür*], for otherwise it could not be imputed. But there cannot be any further cognition of the subjective ground or the cause of this adoption…for otherwise we would have to adduce still another maxim into which the disposition would have to be incorporated, and this maxim must in turn have its ground (6:25).

This passage is important because it makes clear just how unique the exercise of *Willkür* really is. *Willkür* is spontaneous in the sense that it is not guided by an antecedent principle of action; it is strictly speaking lawless. Kant makes this clear in his suggestion that there cannot be any further cognition of why an evil *Gesinnung* is adopted. In order to attain this, we would have to understand the choice as embedded in a yet deeper principle of action, one that—by definition—an original and self-constituting choice does not permit.

The rational opacity of the original choice of one's *Gesinnung* is emphasized again in a passage we have already looked at. Kant writes, "since the fall from good into evil (if we seriously consider that

evil originates from freedom) is no more comprehensible than the ascent from evil back to the good, then the possibility of this last cannot be disputed" (6:45). Bracketing Kant's misleading claim that humans fall from good to evil (Kant thinks we merely have a predisposition to good), Kant claims that the original choice of evil is just as incomprehensible as the subsequent move from evil to good. This, in fact, is offered as evidence that conversion is possible. Kant reasons: if the original fall into evil is incomprehensible but possible, then perhaps conversion from evil to good is similarly possible.

Samuel Loncar has recently argued that Kant's reasoning is flawed insofar as the purported parallels between the fall to evil and the subsequent ascension to good are dis-analogous. In support of this claim, Loncar rightly points out that the initial fall is a movement from innocence (or moral neutrality) to evil, while the conversion to good is a movement that begins from a determinant moral disposition.[19] Against Loncar, though, Kant's aim appears to be a bit more modest. Kant is not claiming that there is a parallel insofar as both events rehearse the same exact movements; he is claiming that the two are parallel insofar as they are both *incomprehensible*, albeit in different ways. Furthermore, Kant appears fully aware that the two movements are distinct in the way that Loncar suggests. If the movement from evil to good was not importantly different than the movement from innocence to evil, Kant would have a ready-made solution to his conversion problem: namely, conversion by spontaneous choice. The very fact that Kant *doesn't* recommend a parallel solution suggests that he understands that the situations aren't parallel in this more specific sense.

So why doesn't Kant simply opt for the same volitional mechanism that explains an agent's original choice? Even if the movement from a morally determinant disposition to another is different than the initial movement from moral neutrality to evil, why not think that *Willkür* is equally suited to both tasks? On this point, I think Loncar diagnoses the issue correctly. He notes that the account of agency that Kant desires is one where action is appropriately rule-governed; as Kant claims in Section III of the *Groundwork*: "What, then, can freedom of the will be other than autonomy, that is, the will's property of being a law to itself" (4:447). However, when it comes to the original exercise of *Willkür*—the agent's choice of evil—there is no law which the agent acts from: he acts neither from the legislative will, nor established character. While this original and lawless act is necessary to explain how an agent comes to have moral character to begin with, Kant is hesitant to appeal to the same mechanism when an agent's moral character is already established. Once an agent has an established moral disposition, his actions are autonomous only insofar as they are performed in accordance with his character. While Kant's interest in indexing action to character makes it even more difficult to understand how an agent's initial act of evil can be comprehensibly attributed to the agent, the motivation for the move is clear enough. Kant is concerned to respond to Leibnizian and Humean worries concerning action attribution: namely, that autonomous action must be rationally or causally linked to some central element of an agent's underlying character.

Section 2.2: Conversion by "Gradual Reformation"

[19] Loncar 2013, 355-6.

If certain Leibnizian-Humean worries pressure Kant to reject a "spontaneous choice" account of conversion, perhaps a "Leibnizian"[20] account is in order. Such an account would show how conversion is a function of making certain implicit character traits explicit. If action from one's character is a necessary condition for action itself, then perhaps we can give an account of agent conversion that shows how certain latent elements of a person's character gain volitional prominence. If so, then we could explain conversion as an offspring of an agent's antecedent dispositions. The conceptual ingredients for such an account are present in Kant's discussion of both the "germ of goodness" and the idea of gradual moral reform.

As we saw in our discussion of the "original predispositions to good," Kant thinks that all human beings have a predisposition to "personality." In the closing remark of Part I, Kant calls this predisposition a "seed [*Keim*] of goodness…a seed that cannot be extirpated or corrupted" (6:45).[21] Sticking with the botanical metaphor that has guided the discussion thus far, one might think that the movement from evil to good consists in the gradual growth of the seed to goodness. On this interpretation, the seed slowly matures as a person's good actions nourish it further and further. Eventually, as the seed continues to be nourished, the seed blossoms into a tree that—low and behold—is good. This "gradual reformation" account avoids the worries we encountered in the spontaneous choice account by showing how the movement from evil to good is, at every step, explainable in terms of an agent's underlying character. Perhaps this is the explanatory route Kant should take.

Though it is true that Kant emphasizes a gradual process of character formation that occurs at the empirical level, he is adamant that empirical striving exercises no causal influence on a person's supreme maxim. In a passage footnoted earlier, Kant writes:

> To look for the temporal origin of free actions as free (as though they were natural effects) is therefore a contradiction; and hence also a contradiction to look for the temporal origin of the moral constitution of the human being, so far as this constitution is considered as contingent, for the constitution here means the ground of the exercise of freedom which (just like the determining ground of the power of choice in general) must be sought in the representations of reason alone (6:40).

In a separate but related point, Kant also emphasizes that evil and goodness are qualitatively distinct making it impossible to transition from one to another by gradual steps. He writes:

> It is a peculiarity of Christian morality to represent the moral good as differing from the moral evil, not as heaven from *earth*, but as heaven from *hell*. This is indeed a figurative representation and, as such, a stirring one, yet not any the less philosophically correct in meaning—For it serves to prevent us from thinking of good and evil, the realm of light and the realm of darkness, as bordering on each other and losing themselves into one another by gradual steps (of greater and lesser brightness); but rather to represent them as separated by an immeasurable gap (6:60 fn.).

---

[20] By "Leibnizian account," I do not mean Leibniz's actual account of conversion, but rather an account that attempts to explain conversion by reference to an agent's antecedent character traits. Leibniz and Hume are identified with such an account, because they notably resist libertarian accounts of action like the one offered in § 2.1.

[21] Translation modified from "germ" to "seed."

Though Kant's first point concerning our inability to postulate causal influence from temporal *Handlungen* to noumenal *Gesinnung* is important, this second point about the qualitative divide between evil and goodness is philosophically primary. For Kant, there cannot be a gradual movement from evil to good because there is no intermediate state between these two poles. This follows from Kant's "ethical rigorism," his claim that an agent's "disposition as regards the moral law is never indifferent (neither good nor bad)" (6:24).

Section 2.3: Conversion by "Grace"

Kant's inability to explain self-conversion leads him to formulate what, at first, appears to be a supernatural account of conversion. He writes: "Granted that some supernatural cooperation is also needed to his becoming good or better…[the evil agent] must *accept* this help (which is no small matter)…in this way alone is it possible that the good be imputed to him, and that he be acknowledged a good human being" (6:44). Lest, however, we suspect that Kant genuinely thinks that divine grace is required for conversion, Kant suggests that 'salvation by grace' is just a stand in for 'self-conversion we know not how.' Kant writes:

> But does not the thesis of the innate corruption of the human being with respect to all that is good stand in direct opposition to this restoration through one's efforts? Of course it does, so far as the comprehensibility of, i.e., our *insight* into, its possibility is concerned, or, for that matter, the possibility of anything that must be represented as an event in time (change) and, to this extent, as necessary according to nature, though its opposite must equally be represented, under moral laws, as possible through freedom; it is not however opposed to the possibility of this restoration. For if the moral law commands that we *ought* to be better human beings now, it inescapably follows that we must be *capable* of being better human beings (6:50).

So how does Kant balance the claim that "we must be capable of being better human beings [through our own efforts]," with the claim that "supernatural cooperation is…needed to become good"? For Kant, it seems the invocation of "supernatural cooperation" is just a way of capturing the idea that self-conversion is a *mystery*: we know that we must be capable of it, but we have no idea how. Even if our empirical actions were sufficient to enact such a change (which they aren't), we wouldn't be capable of knowing how many empirical deeds were requisite. Despite our apparent impotence and ignorance, however, we are permitted to "hope" for a "cooperation," the details of which we need not trouble ourselves with: "For here too the principle holds, 'It is not essential…that every human being know what God does, or has done, for his salvation' " (6:52).

Conclusion

Now that we've seen why Kant cannot endorse any of the three explanatory options—conversion by spontaneous choice, gradual reformation, or grace—it becomes a bit clearer why Kant settles on the above cited "mystery" account of conversion. Kant insists that conversion is attributable to an agent, but the only two mechanisms possible—spontaneous choice and gradual reform—each violate a commitment of his moral psychology. Spontaneous choice accounts are unable to plausibly explain how conversion follows from an agent's guiding normative principle; while gradual reformation accounts fall afoul of Kant's rigorism. Though Kant saw no way forward with any of these positions, each goes on to play a central role in the history of German Idealism. Far from being a niche interest of a small set of religious thinkers, the concept of conversion—along with the

models that Kant explores—plays a central role in the thought of the later idealists. While I do not have space to tell this story here, this essay has—I hope—laid a foundation for its development.

Works Cited

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

Allison, Henry. "On the Very Idea of a Propensity to Evil." *Journal of Value Inquiry* 36.2,3 2002: 337-48.

Ameriks, Karl. "Kant on the Good Will." In *Grundlegung zur Metaphysik der Sitten, Ein kooperativer Kommentar*, ed. by Otfried Höffe, Frankfurt am Main: Vittorio Klostermann, 1989: 45-65.

Anderson-Gold, Sharon. "God and Community: An Inquiry into the Religious Implications of the Highest Good." In *Kant's Philosophy of Religion Reconsidered*, ed. by Phillip J. Rossi and Michael Wreen, Bloomington: Indiana University Press, 1991: 113-131.

Frankfurt, Harry. "Autonomy, Necessity, and Love." In *Necessity, Volition, and Love*, Cambridge: Cambridge University Press, 1999: 129-141.

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." In *The Importance of What We Care About*, Cambridge: Cambridge University Press, 1998: 11-25.

Kant, Immanuel. *Groundwork of The Metaphysics of Morals*. Trans. and ed. by Mary J. Gregor. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press, 1996. 37-108.

Kant, Immanuel. *Religion Within the Boundaries of Mere Reason*. Trans. and ed. by Allen W. Wood. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press, 1996. 39-216.

Kemp, Ryan. "The Contingency of Evil: Rethinking the Problem of Universal Evil in Kant's *Religion*." In *Rethinking Kant: Volume 3*, ed. by Oliver Thorndike, Newcastle: Cambridge Scholars, 2011: 100-123.

Kosch, Michelle. *Freedom and Reason in Kant, Schelling, and Kierkegaard*. Oxford: Oxford University Press, 2006.

Loncar, Samuel. "Converting the Kantian Self: Radical Evil, Agency, and Conversion in Kant's Religion within the Boundaries of Mere Reason." *Kant-Studien* 104.3 2013: 346-366.

Muchnik, Pablo. "An Alternative Proof of the Universal Propensity to Evil." In *Kant's Anatomy of Evil*, ed. by Sharon Anderson-Gold and Pablo Muchnik, Cambridge: Cambridge University Press, 2010: 116-143.

Palmquist, Stephen R. "Kant's Quasi-Transcendental Argument for a Necessary and Universal Evil Propensity in Human Nature." *The Southern Journal of Philosophy* XLVI 2008: 261-97.

Wood, Allen W. "Unsociable sociability: The Anthropological Basis of Kantian Ethics." *Philosophical Topics* 1991: 325-351.