Kant's Quasi-Transcendental Argument for a Necessary and Universal Evil Propensity in Human Natur

Palmquist, Stephen R.

Published in:
Southern Journal of Philosophy

DOI:
10.1111/j.2041-6962.2008.tb00079.x

Published: 01/06/2008

Document Version:
Early version, also known as preprint

Kant's Quasi-Transcendental Argument for a Necessary and Universal Evil Propensity in Human Nature

1. The Necessity of a Formal Proof

In Part One of *Religion within the Bounds of Bare Reason*,[1] Kant refers on several occasions to a "proof" that human nature is exposed at its root to an "evil propensity". The most notorious of these references comes near the beginning of Section III, where he writes (*Religion,* 32-33): "We can spare ourselves the formal proof that there must be such a corrupt propensity in the human being, in view of the multitude of woeful examples that the experience of human *deeds* parades before us." Commentators have typically taken this statement, followed as it is by several examples of human evil in various empirical manifestations, as Kant's excuse for failing to provide a "formal proof" of his central claim in Part One of *Religion*, that an evil propensity is a necessary and universal component of human nature.[2]

If Kant really is confessing that, in view of the overwhelming weight of the empirical evidence that people are evil, he will not bother to construct a formal proof, then the reputation of *Religion*, as being a book that is high on challenging claims but low on persuasive philosophical argumentation, would surely be more than justified. For in Part One (especially Section II) Kant does not merely make the matter-of-fact claim that the propensity of human nature *is* evil; he repeatedly implies and sometimes explicitly states that human nature *must* (or *needs* to) have such a propensity, and that it is therefore *universal*.[3] In view of the "rigorist" stance he takes in the opening (unnumbered) section of Part One (whereby a human disposition must be *either* good *or* evil, never a mixture of the two and never merely neutral [*Religion*, 22-26]), Kant's claim that we inevitably tend toward the evil side of the equation surely cries out for a formal proof – as well as an explanation of how (as Kant also repeatedly insists) each person can be held *responsible* for making moral choices in line with this necessary evil propensity. For if the propensity really is *necessary*, this would seem to compromise the autonomy that Kant takes such pains to uphold throughout his philosophical writings as the prize possession of all human persons. Any

persuasive account of Kant's "proof" of universal human evil must therefore explain how Kant thought a "necessary" propensity could nevertheless be *freely* chosen. His list of empirical examples clearly fails to accomplish this goal, for it neither requires *all* persons to succumb to evil nor shows that those who do choose evil choose it *freely*. Did he really mean to imply that such a list could *replace* a formal proof?

The usual affirmative answer interpreters give to this question not only makes Kant out to be a third-class philosopher, but totally ignores the fact that Kant himself answers this question negatively just a few pages later. Section III ends with a paraphrased quotation of Romans 3:9-10 (*Religion*, 39): "There is no distinction here, they are all under sin – there is none righteous (in the spirit of the law), no, not one." To this affirmation of a biblical principle that was as unpopular among philosophers of Kant's day as it is today, Kant attaches a footnote that begins as follows (*Religion*, 39n):

> 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 – though experience can never expose the root of evil in the supreme maxim of a free power of choice in relation to the law, for, as *intelligible* deed, the maxim precedes all experience.

Obviously, Kant did *not* intend the list of empirical examples of evil deeds in Section III to constitute a proof of evil all on its own, for he here states that the "appropriate proof" (presumably, the same "formal proof" referred to in the previously-quoted passage) can be found in *Section II*. The problem is that nobody up to now (as far as I am aware) has ever actually *located* the proof Kant refers to in this footnote. Those commentators who show an awareness of this problem typically assume Kant never actually supplied the proof in question – despite his claim in the above-quoted footnote to have done so.

Allen Wood, for example, thinks this "appropriate proof" refers merely to the definition of evil Kant provides in Section II, as constituting the reversal of the proper order in the incentives a person adopts when formulating a maxim for action.[4] Wood calls attention to Kant's claim, in the sentence just before the start of Section I, that the attribution of an evil propensity to the entire human race, without exception, "can only be demonstrated later on, if it transpires from anthropological research that the grounds that justify us in attributing one of these characters [either good or evil] to a human being as

innate are of such a nature that there is no cause for exempting anyone from it."[5] He thinks this passage shows that Kant never intends to determine, a priori, whether the human species as a whole has an evil propensity; although some individuals clearly are evil, the question of whether this applies to *all* human beings is a matter for anthropology to decide.[6] While in his earliest work Wood portrayed the evil propensity as an "empirical generalization", he now regards that view as "naïve" and puts in its place an appeal to Kantian anthropology.[7] Along the lines argued by Sharon Anderson-Gold, Wood now associates radical evil with "the human trait of unsociable sociability."[8] The latter is undoubtedly bound up with Kant's conception of evil; however, it relates not to any *proof* of evil but to Kant's arguments for the *church* as a necessary social structure for combating the empirical manifestation of the evil propensity. As we shall see (especially in §4), Wood's conflation of this theory from Part Three with the need for a *proof* of the evil propensity[9] conflicts with the most natural way of reading numerous statements in Part One, where Kant does appear to be arguing that human nature *as such* has (and *must* have!) an evil propensity.

Given that Kant's emphasis in Sections I and II is *exclusively* on the claims that the human predisposition is good, while its propensity is evil, Kant's appeal to anthropological research in *Religion*, 25, could be read more naturally as a concession to any religious person who wishes to claim that at least *one* historical human being (e.g., Jesus of Nazareth) displayed no empirical evidence of having an evil propensity. The point of the passage, after all, is to raise the question of whether some human being(s) might be an *exception* to the general rule of the universality of evil in the human species. Kant deals with precisely this possibility in Part Two, Section One, Subsection B (*Religion*, 63-66), where he discusses how a philosophically-responsible believer might conceive of such a holy human being's fundamental character. The "anthropological research" Kant had in mind (if I am right) would be directed at the life and teachings of Jesus (or any other historical person who might be held up by religious people as an ideal of moral perfection) – a research paradigm that really did become a central focus of post-Kantian theologians, in the form of the "quest for the historical Jesus". An interesting question is just how such anthropological research could justify an exception being made to a formal proof that human nature in general has a propensity to evil. But this question is beyond the scope of the present essay, where our concern is to locate (if possible) and understand the structure of the formal proof itself.

Another possible response to Kant's various references to the need for a "proof" of the evil propensity in human nature is adopted by Michalson, who portrays Kant's whole religious philosophy as producing a "series of wobbles" between fundamentally incompatible commitments – such as those of orthodox Christian theology and Newtonian mechanistic (and ultimately atheistic, or at least agnostic) science, or more troublingly, conceptual inconsistencies Michalson sees within Kant's own claims – leaving his whole "religious philosophy rippling with instability."[10] If commentators such as Michalson are correct, then we can simply forego any attempt to harmonize Kant's various references to a "proof" of the evil propensity, because the whole topic constitutes but a typical example of the fundamental "instability" of Kant's theory of religion. By showing the dire consequences for Kant's theory if (or, indeed, *because*) up to now nobody has found the proof Kant claimed he gave, Michalson's approach highlights (negatively) how important it would be if that proof *could* be identified. For instead of participating in the search for a proof, Michalson projects this failure onto Kant's text, assuming Kant was merely confused.

Later in this essay (especially §4) I shall attempt to explain exactly what Kant meant by these (and other) references to a proof that an evil propensity *must* lie at the very root of human nature. For now, however, it will suffice to conclude these initial reflections by pointing out that, in order to harmonize the above-quoted passages, we must assume that Section II does not contain the *whole* formal proof; for Kant tells us in the quoted footnote that Section III's examples play the role of "corroborating" the formal (presumably a priori) judgment presented in Section II. In order for such corroboration to be possible, Kant's earlier statement, "We can spare ourselves the formal proof", must not mean that he really intends his examples to *stand in place* of a proper (synthetic a priori) proof, but rather that the empirical evidence of evil in human nature is so plentiful that *nobody can seriously doubt* the validity of the basic claim. That is, the role of Section III in the overall argument of Part One is to confirm (as if it were ever in doubt!) that *evil really does exist*. We can therefore read the earlier passage as claiming that the empirical evidence is so compelling that it *can* persuade us even without being part of a formal proof, *even though for the Critical philosopher such examples do not constitute a sufficient proof*. To say Section III is not a sufficient proof *on its own* does not imply, however, that the citation of such examples is not *necessitated*

in order to corroborate the "appropriate proof" Kant claims to have provided in Section II. My goal in this essay is to identify the exact role played by these two sections in Kant's attempt to prove that the human propensity *must* be evil – a goal that will be fulfilled in §4. Before examining the precise details of that relationship, let us first assess (in §2) the attempts other scholars have made to identify a "formal proof" that would be "appropriate" to fulfill this tall task. Against the backdrop of this on-going debate, we can then establish (in §3) exactly what *type* of proof we should expect to find.

2. Attempts to Reconstruct Kant's Allegedly Missing Proof

In contrast to Wood and Michalson, several commentators who have recognized the *need* for a formal proof in Kant's argument, but who share the conviction that Kant himself does not provide it for us in any complete or persuasive form, have attempted instead to reconstruct a proof that *could* fulfill the task Kant seems to have carved out for such a proof. In this section I shall examine three prominent examples of such a reconstructed proof, as proposed by Henry Allison, Seiriol Morgan, and Peter Fenves.

Henry Allison adopts an approach to the problem of the missing proof in Part One of Kant's *Religion* that is more amenable to the spirit of Critical philosophy (and certainly more affirmative of Kant) than approaches such as Michalson's. According to Allison,[11] Kant portrays the evil propensity in human nature "as a postulate of morally practical reason and, therefore, as a synthetic a priori claim."[12] Like Wood and Michalson, he assumes an important proof is missing in Kant's text: anything synthetic a priori (i.e., anything *transcendental*) must be supported by a purely philosophical proof, often called a "deduction",[13] yet *nowhere* in *Religion* does Kant explicitly refer to a deduction of evil.[14] If, as seems likely, the "appropriate proof" of evil in Part One is supposed to take the form of the type of argument Kant himself regarded as supremely "appropriate" for philosophers – sometimes called, more generally, a "transcendental argument" – then, as Allison points out, Kant appears to have failed rather blatantly in the very task he explicitly claims (in the footnote quoted in §1) to have *fulfilled*.

Allison responds to this unfortunate situation by seeking to lend Kant a helping hand: he reconstructs on Kant's behalf a transcendental deduction for the evil propensity in human nature. After demonstrating (in stark contrast to Michalson's later work) that Kant's concept of radical evil is

thoroughly consistent with features of his moral philosophy that were already well developed in the *Groundwork,*[15] Allison argues that Kant's main task in Section II is to justify the "hybrid notion of a freely chosen propensity".[16] He does this through "a distinction between two meanings of the term 'act' (*That*)": "intelligible action" and "sensible action".[17] The evil propensity is thereby identified as an intelligible "act", whereby a person adopts "a *deliberative* tendency … to allow … nonmoral considerations stemming from inclination to outweigh moral ones."[18] Allison rightly observes the "peculiar" nature of this alleged "act", inasmuch as Kant does not seem to think this underlying ground of all empirical action is "explicitly and self-consciously adopted by an agent."[19] Rather, Kant's claim seems to be that for this reason, the act of giving in to the evil propensity "must be presupposed by, rather than revealed in, moral reflection."[20] As we shall see in §4, this important insight is the key to the "appropriate proof" that Kant believed he had presented in Section II. For near the end of that section Kant says the "intelligible deed" of giving in to the evil propensity is "cognizable through reason alone apart from any temporal condition" (*Religion*, 31). Oddly, after quoting this statement, Allison says "the assumption of such a propensity functions as a postulate of morally practical reason";[21] yet the form of Kant's argument, as we shall see, is nothing like that of the practical postulates in the second *Critique*.

The practical postulates in the Dialectic of Kant's *Critique of Practical Reason* come in at the end of Kant's exposition, as a rational means of saving morality (virtuous action) from the charge of being meaningless, given the lack of proportion we experience in this life between virtue and happiness. By contrast, Kant's argument regarding the propensity to evil comes in at the very outset of his exposition in *Religion*, as a way of setting the basic parameters for the entire book. It is, as Allison rightly states near the beginning of his account yet fails to flesh out, transcendental (synthetic a priori) in form; as such, the evil propensity (together with the good predisposition) has the same function for *Religion* as space and time have for the first *Critique*, and as the fact of freedom has for the second.

As one of those commentators who regards Kant's appeal to examples in Section III as a cop-out (see note 2, above), Allison criticizes Kant for treating his assumption of the evil propensity as if it were "an unproblematic empirical generalization."[22] In an effort to fill the gap he thinks Kant has left by this otherwise unforgivable blunder, Allison proposes a new "key" to the deduction Kant *should* have

presented: the deduction must begin by determining "what a propensity to good would be like if, *per impossibile*, a human being could possess one."[23] Allison admits that "Kant does not describe it in so many words"; yet he believes that for Kant a good propensity would mean "that for an agent the moral incentive would, as a matter of course, always outweigh the incentive of self-love."[24] For such a person, "the law would not take the form of an imperative and moral requirements would not be viewed as duties."[25] This amounts to what Kant elsewhere calls "holiness", an option he thinks is not open to human moral agents; for "we are never beyond the possibility of temptation and the need for moral constraint."[26]

Allison sees two ways of applying this point about the holiness of a propensity to good to defend Kant's claim in Section II that the human propensity must be evil. The first option is to appeal to Kant's rigorism: if human nature must be *either* good *or* evil, if holiness is foreign to our nature, and if a good propensity would amount to holiness, then "the inference to a universal propensity to evil follows" as a matter of deductive certainty.[27] However, Allison rightly points out that this reconstruction of Kant's argument "only succeeds in trivializing it."[28] Indeed, it reduces Kant's argument to a merely analytic claim about the nature of the concepts of good and evil. Allison believes the second option is more promising: we can "put some bite back into the doctrine that there is a universal propensity to evil" by focusing on the "fact that we only obey the law reluctantly", for this implies that we do have "a tendency to let ourselves be tempted or 'induced' by inclinations to violate the moral law even while recognizing its authority."[29] Allison thinks this suffices as a deduction of Kant's claim that the human propensity is universally evil, yet without compromising the responsibility of the individual moral agent. But the "fact" he refers to is *empirical*, a matter of human moral psychology, so it cannot possibly establish a *transcendental* conclusion. Allison's attempted reconstruction of Kant's argument therefore leaves us with two equally undesirable options: a trivial but certain truth, or a weighty but merely empirical (and therefore uncertain) truth. If Kant believed his own argument was philosophically "appropriate" (i.e., synthetic a priori), then surely neither of Allison's two options suffices to show us its true form.

Along similar lines to Allison, a recent article by Seiriol Morgan attempts to fill the alleged gap in Kant's exposition by reconstructing an argument that Kant himself *should* (or at least *could*) have presented, but did not. According to Morgan, "the missing formal proof of humanity's radical evil" can be

established through a very different type of reconstruction than the procedure adopted by Allison. Morgan agrees with Allison's claim that the first *Critique* requires a deduction to justify any alleged synthetic a priori proposition and that in Part One of *Religion* Kant's "claims are crying out for a transcendental deduction that he does not provide."[30] Unlike Allison, however, Morgan thinks "Kant's execution" of a proof of evil "is self-contradictory", even though "a synthetic a priori argument for a universal human propensity to evil is in fact available"; the reconstructive interpreter must therefore be willing to abandon certain parts of Kant's theory.[31] The heart of Kant's alleged self-contradiction is his claim "that we necessarily freely choose [evil] rather than [good], whereas freedom and necessity surely exclude one another."[32] Whereas Morgan, like Michalson and many other commentators, assumes Kant's position here "is a muddle", producing an "internally inconsistent" position,[33] I shall argue in §4 that in order to understand Kant's position properly, we must *embrace* this very paradox, that we *all* (as a species) *inevitably* choose an evil disposition, yet that this choice is entirely free and therefore carries with it moral responsibility. Kant means just what he says. For his strategy in *Religion*, as we shall see, is to show how this very paradox is what makes religion itself both possible and necessary for all human societies.

Having rejected Kant's own argument as self-contradictory (as a result of misconstruing the nature of the argument Kant actually advances[34]), Morgan sees no option other than reconstructing Kant's argument by doing away with the paradox. (Morgan shows no awareness that for Kant this procedure, if successful, would also have the effect of doing away with religion [see §4].) After a lengthy exposition on the basic elements of Kant's moral theory as developed in *Groundwork*, Morgan sets out "to produce the transcendental argument [Kant] alludes to but omits to provide."[35] His argument proceeds as follows. An evil will is one that refuses to recognize any internal constraint on willing (i.e., one that neglects the moral law), and thus (wrongly) interprets its "outer freedom" as "the absence of any and all restraints upon its willing."[36] Morgan claims "it is entirely appropriate to think of the propensity to evil as just this incentive to embrace unrestrained license."[37] For "the will's overabundant affirmation of its own freedom will mean, in practice, the subordination of morality to self-love."[38] Because this argument requires the agent to be "transcendentally free", and because "all [its] steps are a priori", Morgan concludes "the argument I have provided is the formal one we have been looking for."[39]

While Morgan's reconstructed argument of the alleged "missing proof" in Part One of *Religion* may be consistent with the arguments Kant himself advances, it has two weaknesses. First, by rejecting at the outset the very paradox Kant so clearly affirms, Morgan stands in danger of overlooking what might end up *constituting* the basic form of a proof that Kant does, in fact, present in Part One. This is a possibility I shall explore further in §4. Second, Morgan *claims* his "argument also establishes the universality of the propensity";[40] yet he never demonstrates how it does so. Morgan's argument is formal and a priori, but only in the same *analytic* sense as Allison's first option: he ingeniously presents a definition of "outer freedom" as self-deceptive (i.e., as excluding the inner freedom of the moral law) and claims (perhaps rightly) that this is an assumption adopted by people who have succumbed to the evil propensity. But how does this clarification of the *concept* amount to a transcendental proof?

Morgan's account does elucidate some important aspects of Kant's theory, as he explains in the second half of his article.[41] Perhaps most significant is that his account of Kantian evil explains how the *incentive* to be evil "bubbles up from freedom itself, and since it emerges from this source it cannot be caused" (92) – the latter being the danger with any account of evil that appeals specifically to temptations as originating in our inclinations. However, notwithstanding the merits of this significant insight, Morgan has not presented any argument that reason's self-deceiving assumption is *necessary* and/or *universal*; the only time he comes close to doing so is in a paragraph that is not actually part of his "reconstruction", where he merely quotes texts straight from Part One.[42] Surely this further proof of universality and necessity must be explicitly stated, if the argument is to be in any meaningful sense transcendental. Yet Morgan never says either what his reconstructed *transcendental* argument actually *is*, or *why* commentators have been unable to locate it for over two centuries. Any persuasive explanation of Kant's proof of evil must account for both of these conundrums.

The suggestions of Allison and Morgan are interesting and creative attempts to make up for the glaring shortcoming they believe Kant left in Part One of *Religion*. Such interpreters are forced to regard the passage quoted at the beginning of this essay as nothing but a lame excuse for Kant's utter failure to provide the transcendental argument he *knew* would be needed in order to demonstrate that an evil propensity must exist (universally) in human nature.[43] Yet if this were Kant's intention in quoting such

examples, then his strategy was not just lame but foolish as well, because the fact that the many empirical examples make the reality of evil obvious only *intensifies* the need for a formal proof; it cannot stand in place of it! Did Kant really fail to grasp such a basic fact about what his own theory requires? Or could he perhaps have been saying something quite different in the passage quoted at the outset? In particular, could he have been claiming he *had* succeeded in giving a formal proof, even though the examples he was about to cite would suffice to persuade most readers (i.e., readers who, unlike Kant himself, do not require a transcendental argument)? The fact that the string of examples Kant provides in Section III is followed by a rehashing of some basic definitions and descriptions that he had already set forth with sufficient clarity in earlier sections could be taken as evidence that Kant really did think his "appropriate proof" was somehow *completed* by these empirical examples. I shall explore this possibility further in §4, below.

As we saw in §1, Kant's footnote at the end of Section III clearly alleges that the "missing proof" Allison and Morgan try to reconstruct is *already present*, in Section II; yet Kant's interpreters have found nothing like a transcendental argument for evil in that section. We saw that interpreters such as Michalson have concluded on this basis that Kant was *at odds with himself* over this issue, supposedly not knowing what he really believed or wished to argue. Wood thinks Allison's "implausible" reconstruction is "especially unpromising", inasmuch as the mere fact of finitude does not (as Allison wrongly assumes) analytically imply that the human will is unholy.[44] But even if it did carry this implication, such an argument would (as Wood notes) be "trivial", because it would merely explain either what it means, or how we can be motivated, to make evil choices; it would not confirm the more substantive claim that the evil propensity is a necessary and universal component of human nature, this alone being the claim that Kant (as quoted in §1) thought stands in need of an "appropriate proof" for its justification.[45]

Morgan says Wood's "reductionist interpretation is a reading of last resort, since it threatens to make the phenomena Wood focuses on the unfortunate consequence of a natural process".[46] Ironically, however, Morgan's expressed desire "to tone down somewhat [Kant's] indictment of the human race"[47] ends up making room for precisely the view Wood defends (see §1). We are, says Morgan, only "*drawn* toward evil by the will's inner yearning for limitless self-assertion…. But that the illicit exerts an inevitable pull on all human beings does not of course entail that everyone embraces evil as their

fundamental commitment".[48] Just as we found in assessing Wood's interpretation, so also Morgan's position need not compel us "to tone down" Kant, nor is his position necessarily incompatible with Wood's (see note 45, above). For none of these even comes close to identifying what Kant had in mind when he claimed he had, in fact, presented an "appropriate proof" for evil in Section II.

Before considering a new way to solve this long-standing interpretive problem, let us examine one further attempt to fill the apparent gap in Kant's exposition: Peter Fenves offers an interesting post-modernist angle on this problem that ends up defending a position strikingly similar to that of Allison and Morgan. Playing on the metaphor of reason's self-exposure, as implied by Kant's use of "*bloßen*"[49] in the title and elsewhere throughout the book, Fenves claims that in *Religion* Kant exhibits two conflicting tendencies: consciously, "Kant emphasizes that *Religion* makes no attempt to conceal anything", yet everywhere one turns throughout the book, the reader encounters an apparently unconscious "secrecy" – most notably in Kant's claim that "the ground of … freedom … is 'noumenal'" and therefore unknowable.[50] On the one hand, Fenves quotes Kant's appeal to "anthropological research" (*Religion*, 25), claiming (like Wood) that Kant regards this "as a final court of appeal" for his theory of evil.[51] With typical post-modern playfulness, he wryly suggests the term "radical evil" means human beings are *only* corrupt at the root, "not altogether evil, corrupt root and branch."[52] On the other hand, Fenves points to Kant's "promise" to "deliver a 'proper proof' of the thesis of radical evil" and claims he could keep such a promise only by obtaining definite knowledge of the noumenal world, but that this would transcend the bounds of knowledge set by the first *Critique*. This, according to Fenves, is the real reason "Kant does not demonstrate anything, least of all the thesis of radical evil", for he is making claims that he himself cannot fulfill.[53] To *Religion*, 39n, he replies: "By declaring that the 'proof proper' has been demonstrated in the previous section without having conducted anything but 'improper' methods of proofs, Kant reveals the secret in secret, and thus does not properly reveal it but does not properly conceal it either."[54]

Instead of blindly trusting Kant in his theory that evil is necessary in human nature, Fenves points out that "the very thesis … is reason for universal distrust"[55] – an interpretation that neatly (though secretly, for Fenves never admits this openly) portrays Kant as the *father* of post-modernism! As such, he suggests that the ironic condition for "understanding the thesis of radical evil" is Kant's (secret!) demand

"that the 'subjective necessity of evil' be recognized in absence of a proper proof of its objective foundation".[56] Perhaps the ultimate irony of Fenves' interpretation is that, if he is right about the deep ironies in Kant's theory of evil, then Kant *does* actually end up providing something like a proof of evil. For the "hiding" of one's own responsibility from oneself becomes the very essence of the propensity to evil on this reading of Kant. If "all moral judgment becomes clouded by clear conscience" and if Kant "makes the confession of untrustworthiness mandatory for all those who want to exonerate themselves of the accusation that they are untrustworthy" (87), then even though Fenves *claims* he has shown that Kant's true conclusion is that we are all "entirely uncertain" about our moral status,[57] in fact he has shown that Kant (paradoxically) *demonstrates* (at least by example) that the human propensity must be evil!

As we have seen, interpreters of Kant's *Religion* have not hesitated to affirm the main two points of Fenves' rather controversial reading of Part One: that Kant openly claims a proof of evil is necessary and yet, if he really did construct such a proof, he hid it so well that nobody has been able to find it.[58] Yet according to Kant's own claims, nothing less is at stake here than the future (or at least, the rationality) of religion itself. Clarifying the precise nature and status of the human propensity is essential to a proper understanding of religion. For if it turns out *either* that human nature is not exposed to an evil propensity but rather starts out with a propensity toward the good,[59] *or* that human beings start out with an evil predisposition so *inevitable* that no individual person can be held responsible for it, then the *problem* that gives rise to the need for religion will have evaporated. The problem is that, by breaking the moral law, we bring guilt upon ourselves, guilt so radical that it destroys our ability to fulfill the very purpose of human life (i.e., to obey the moral law). Religion arises in every human culture precisely because this problem inevitably cries out for a solution; if it can be solved as easily as either denying we are *evil* or denying we are *responsible* for evil, then the rationality of the claim that human society *needs* religion will also be called into question.

My strategy in the remainder of this article is twofold. In §3 we shall see that the various references to "proof" and "proving" throughout Part One suggest the required proof should take the typical form of a *transcendental* argument, although Kant avoids using this technical term. An overview of the structure of a typical transcendental argument will therefore give us the basis for the crucial clue to

the solution of this puzzle. In §4 I shall argue that, if we attend to the architectonic relation between the four sections of Part One, viewing them as components of a single, systematic argument, then the basic structure of the "formal proof" that has defied detection for so long will become unmistakably clear.

3. Identifying the Structure of an "Appropriate" (Transcendental?) Argument

Having noted in §2 the inadequacy of previous attempts to identify a "formal proof" that would confirm the necessary and universal status Kant claims for the evil propensity in human nature, and in light of Kant's explicit claim that he has already provided an "appropriate proof" (see §1), I propose we take a step back and ask just what we should expect the elusive proof to *look* like, were we actually to find it in the text of Part One, as he claimed. As we have seen, Allison assumes the proof should take the form of a *deduction*, a type of proof Kant employs whenever he needs to justify the objective reality of a *pure concept* for use by the understanding.[60] (I have elsewhere referred to this use as adopting the "logical perspective".[61]) However, in Part One of *Religion* Kant adopts not the logical but the *transcendental* perspective: he seeks to establish the fundamental boundary conditions that must be assumed in order to explain how religion is *possible* at all.[62] Whereas Allison was wrong to search for a deduction as such, a more general search for a *transcendental argument* (or an argument that performs a similar function) would seem very appropriate.

The parenthetical qualification in the foregoing sentence is necessitated by the fact that Kant himself does not actually use the term "transcendental" anywhere in *Religion*. One might pass this off as a result of his desire to make the book's contents comprehensible to non-philosophers – a goal he alludes to at the end of the second edition Preface (*Religion*, 14). However, the problem with attributing a "transcendental" status to any argument in *Religion* runs deeper than this. For as Konrad Cramer points out, Kant defines "transcendental philosophy" in the Introduction to the first *Critique* in a way that appears to make it exclusively theoretical (or epistemological), standing in sharp contrast to its practical complement, moral philosophy. After examining a wide variety of relevant texts on the nature of transcendental philosophy, with special emphasis on the Introduction to the first *Critique*, Cramer concludes that a strict application of Kant's distinction "is untenable" ("*ist unhaltbar*") since the pure part

of moral philosophy also meets the basic criterion of appealing to "no concept of empirical *origin*" ("*kein Begriff empirishcen* Ursprungs").[63] Any argument that seeks to justify the validity of such non-empirical concepts would seem to qualify as potentially "transcendental", even though Kant himself tends to reserve the use of this technical term for arguments employed in his *theoretical* philosophy. Although we therefore should not expect to find Kant referring explicitly to a "transcendental" argument in *Religion*, we may well find him using the transcendental *approach* to justifying the claim that an evil propensity is a necessary and universal aspect of human nature. For he clearly and repeatedly insists throughout Part One that the human propensity is not something empirical, but is just the sort of concept with a non-empirical *origin* that (as Cramer points out) is the focus of all genuinely transcendental philosophy.

What, then, *is* a transcendental argument? This is not the place for a thorough review of the immense literature on this topic that emerged during the second half of the twentieth century in the wake of Peter Strawson's interpretation and Barry Stroud's controversial article on the subject.[64] Instead, I shall briefly summarize one representative approach to describing the typical form this type of argument is generally expected to take. I shall then show in §4 that in Part One of *Religion* Kant *does* present an argument that corresponds exactly to this standard form. However, in view of Cramer's demonstration that Kant preferred to reserve the word "transcendental" for explicitly theoretical applications, I shall distinguish Kant's "appropriate proof" in Part One from a *technically* transcendental argument (while at the same time acknowledging that Kant himself never employs this term to describe his argument for the evil propensity in human nature), by referring to Kant's proof as a "*quasi*-transcendental argument." Before returning to the text of Part One in search of such an argument, let us clarify the form transcendental arguments in general typically take.

Ralph Walker, in his book, *Kant*, provides one of the clearest accounts of the basic form of Kant's transcendental arguments.[65] After explaining how Strawson and others employ Kantian-style transcendental arguments as a powerful way of responding to the skeptic, Walker suggests all such arguments follow the same basic structure. The purpose of the argument is to persuade the skeptic who doubts the validity of some non-empirical (i.e., synthetic a priori) concept, *p*, that the concept *must* be true, otherwise something the skeptic does not doubt would not even be *possible*. What the skeptic in the realm

of theoretical knowledge does not doubt is the initial givenness of "experience in general" – or, as Walker suggests (going beyond Kant's explicit statements), "intelligible thought".[66] Any skeptic who admits the reality of this "given" will be compelled by the transcendental argument either to admit that *p* is valid or to give up the cherished notion that our experience is genuine (or intelligible). After arguing that the crucial second step in such arguments must be *analytic* (for if it is synthetic a priori, it will not persuade the Humean skeptic, who only accepts empirical and analytic truths),[67] Walker distills them into the following basic three-step form:[68]

> We have experience …
>
> It is analytic that the truth of *p* is a necessary condition for experience ...
>
> Therefore, *p*.

In the first *Critique*, Kant employs this type of argument to prove that space, time, and the categories are necessary and universal conditions for the possibility of experience. If his arguments hold, then anyone who rejects the validity of any one of these synthetic a priori elements would essentially be denying the very possibility of experience itself – an outcome that does not settle well even with the skeptic.

Strictly speaking (as Cramer points out), such arguments apply only to Kant's theoretical philosophy, where "experience in general" is the subject-matter under discussion. However, commentators generally agree that Kant employs arguments with the same basic structure throughout his Critical writings.[69] Of course, the first premise of the argument will have to change when the standpoint (i.e., subject-matter) changes. For example, moral skeptics will be persuaded by the similarly structured arguments in the second *Critique* only if they admit as a starting-point the premise that we do have moral experience (e.g., the experience of free choice). Likewise, in *Religion* Kant appears to be taking as a "given" the reality of religious beliefs and practices in human life that cries out for a rational explanation.

If I am right in claiming that for Kant an "appropriate proof" that the human propensity is evil would take the form of a quasi-transcendental argument, then the first premise of such an argument (following Walker's suggested structure) would have to be something like: "We have religious experience" (not necessarily in any mystical sense, but in the sense of being compelled to adopt beliefs and practices promoted by a religious tradition). Of course, this premise is far more controversial than the

premise of the standard transcendental argument. Nevertheless, Kant apparently viewed the typical reader of *Religion* as someone who would *accept* this premise. In the first paragraph of the first edition Preface, he refers to the "need" a person may have for "the idea of another being above him in order to recognize his duty" (3) and/or for "an incentive other than the law itself." Kant's main point in that opening paragraph is that this need *cannot* be accounted for by an appeal to morality, for "on its own behalf morality in no way needs religion" (3). But in the often neglected second paragraph, Kant goes on to say, *from another standpoint*, that such needs may arise "as a necessary consequence" of our moral nature (4). As such, he openly states as the *given* of his inquiry (6): "the idea of a mighty moral lawgiver outside the human being, in whose will the ultimate end (of the creation of the world) is what can and at the same [time] ought to be the ultimate human end." The reason he can make this assumption is made clear in the second half of the Preface (7-11), where he calls out to *the biblical theologian* (presumably, his primary intended reader) with a plea to take seriously the possibility that philosophy may have something of value to offer, to the extent of even suggesting that theology students (most of whom would be preparing to be *pastors*) should be required to take "a special course on the pure *philosophical* doctrine of religion."[70]

The "skeptic" Kant wants to persuade in *Religion* is obviously not the Humean, who denies the existence of synthetic a priori knowledge, but *theologians* and *pastors* (as well as theology students and perhaps also educated religious believers) who are doubtful about the importance of reflecting philosophically on religion. The reality of religious "experience" (i.e., the rituals and beliefs that are taken to be meaningful within a specific religious tradition) is not called into question, so (as in all successful transcendental arguments) Kant may freely assume it as the basis for his proofs. This is the reason – so often misunderstood by interpreters such as Michalson, who blame their own confusion on Kant's alleged inability to take sides – Kant feels free throughout *Religion* to *assume* the reader will not question the religious concepts and practices he uses as the basis for virtually all of his arguments.

If my argument so far is correct, if the "appropriate proof" Kant claims he has provided is in fact a quasi-transcendental argument aimed at philosophically skeptical religious believers (or anyone who admits that morality does point beyond itself to religious ideas we have a real *need* to adopt), then the crucial second step in his argument in Part One (what he refers to as the "formal proof") should take the

following form: "The truth of *p* is a necessary condition for religious experience."[71] Does Kant ever present an argument that comes close to exhibiting this form? Answering this question, and exploring the various implications an affirmative answer would have, will be my concluding task in this essay.

4. The Architectonic Form of Kant's Quasi-Transcendental Proof

Kant's first hint as to the procedure he will follow in attempting to prove that the human propensity is (and *must* be) evil comes on the second page of Part One (*Religion*, 20) and is worth quoting at length:

> We call a person evil … not because he performs actions that are evil …, but because these are so constituted that they allow the inference of evil maxims in him. Now through experience we can indeed notice unlawful actions, and also notice (at least within ourselves) that they are consciously contrary to law. But we cannot observe maxims, we cannot do so unproblematically even within ourselves; hence the judgment that an agent is an evil human being cannot reliably be based on experience. In order, then, to call a human being evil, it must be possible to infer *a priori* from a number of consciously evil actions, or even from a single one, an underlying evil maxim, and, from this, the presence in the subject of a common ground, itself a maxim, of all particular morally evil maxims.

To read this passage as proposing a "standard of proof" is, Wood claims, both "highly demanding" and "wildly implausible";[72] it makes "the doctrine of radical evil … impossible to argue for", given that my guilt can surely be determined only "from an empirical examination of my conduct. How could anyone be entitled to infer *my* guilty conduct on the basis of general principles about human nature (whether *a priori* or empirical)?"[73] Given his pessimism regarding the plausibility of such an approach, Wood rightly prefers to read the argument Kant is prefiguring as being not about "justifying a verdict of guilty" for specific crimes, but about determining "what people … are disposed to do".[74] Yet he remains concerned about how Kant claims to prove this: "How could it be possible to infer *a priori* from a *single* action an underlying ground of all morally evil maxims (not only for the agent but for the entire human species)?"[75]

If what Kant has in mind in Part One is the construction of a quasi-transcendental argument along

the lines sketched in §3, then the prospects of constructing such a proof in just the way Wood indicates (provided we make the correction suggested in note 73, above) should cease to seem so incredible. For Kant's claim that the human propensity is universally and necessarily evil would then be no less plausible (at least, in principle) than his claim in the first *Critique* that the principle of causality (or the apriority of space and time as pure intuitions) are necessary conditions for the possibility of experience in general. In both cases, his argument moves (just as Walker portrayed for standard transcendental arguments) from the agreed first premise, that the type of experience under consideration does exist, to the inference that a specific a priori principle (or feature of the world) is *necessary* and *universal* for the very possibility of such experience. In both cases, if the skeptic cannot find fault with the "formal" part of the argument (what Walker depicts as the second premise), then the conclusion (that the principle or feature in question is transcendental, forming the very "bounds" [*Grenzen*] of the possibility of the type of experience in question[76]) has been established, QED.

The crucial question, then, is: *Does* Kant, in fact, provide such an argument, somewhere in Part One of *Religion*? Yes, he does. His argument has gone undetected all this time for at least four reasons, presented here in the opposite order of their significance. First, Kant complicates matters slightly by advancing *two* quasi-transcendental arguments in Part One: in Section I, he defends the claim that a *good predisposition* must be inferred to exist in human nature, while in Section II he defends the apparently opposing claim that an *evil propensity* must also hold true for us. In both cases, the form of his argument is quasi-transcendental: in order to account for the possibility (and so also, for religious believers, the reality) of religious ideas (such as God, sin, grace, etc.) we *must* assume these two fundamental building-blocks of the human condition. Just as space and time are different (but complementary) aspects of the synthetic a priori grounding of pure intuition that makes theoretical knowledge possible, so also an originally good potential together with a corruption of that potential at its very "root" must be regarded as conditions whose absence would render the very *origin* of religion impossible to account for. Readers who are either over-hasty or (like Michalson) too quick to blame Kant for their own lack of insight into the meaning of the text tend to confuse Kant's use of "predisposition" and "propensity". But no careful reader should make this mistake, so this first reason on its own would surely not have led to the long

history of blindness commentators have experienced as they have looked *directly at* the quasi-transcendental argument Kant openly tells us he has presented, but without noticing it.

A second possible explanation for the failure of commentators to recognize the quasi-transcendental structure of Kant's argument in Part One is that he presents the two premises in the reverse order of that proposed by Walker. Walker's account, of course, is itself a reconstruction and is not in any sense meant to constrain the way any specific transcendental argument must be presented. But anyone familiar with his account who then opens the pages of *Religion* looking for such an argument in Part One might be misled by the fact that Kant starts with the inference to *p* (the claim that the human propensity is evil) and only then goes on to confirm the premise grounded in experience (that evil does exist). But once again, the attentive reader should not be fooled even by this subtle transposition of premises.

A problem more likely to be responsible for the fact that nobody has yet identified Kant's quasi-transcendental argument is that Kant himself never confirms its status as transcendental. Indeed, as noted above, he never uses the *word* anywhere in *Religion* and even gives the impression in the second edition Preface that the reader need not be versed in the details of Critical philosophy. Instead of telling us plainly how his argument will be structured, he provides only scattered hints, stating his proof will be "a priori", "formal", "appropriate" – and the like. That Kant is sometimes his own worst enemy, thanks to the lack of a clear and consistent mode of exposition, is nowhere better exemplified than in Part One of *Religion*. Even if I succeed in persuading some readers that Kant *does* supply a quasi-transcendental argument for human evil, I will not have (nor attempted to have) shown that it is clearly presented or easy to see. Rather, its obscurity makes the failure of past commentators to notice the argument less surprising.

The major problem preventing past commentators from identifying Kant's proof that human nature has a necessary and universal propensity to evil is that his self-confessed preference for *architectonic reasoning* is typically misunderstood and ignored (if not openly ridiculed) by his readers. Here I cannot demonstrate the full implications of this far-reaching problem, but must simply refer the reader to my previous account of the architectonic form of Kant's approach to philosophizing.[77] In a nutshell, he believed that the Copernican assumption that stands as the over-arching hypothesis of his entire philosophical System (i.e., the assumption that what is philosophically relevant to any experience

will be what the subject reads into it, rather than what we read out from the objective features of our experience) *requires* philosophers to structure their arguments in a manner that is consistent with their pre-determined conception of its proper form. Now, if Kant regarded *Religion* as a genuine component of his philosophical System (as I have argued elsewhere [see note 77]), then this text, too, should show the influence of his architectonic methodology. Indeed it does. For the basic structure of Kant's quasi-transcendental argument for evil can be easily identified in the *outline* (i.e., the *section titles*) of Part One – a fact that is *bound* to be overlooked by anyone who downplays the significance of the architectonic.

A quick look at the headings of the four main sections of Part One confirms that an argument with a quasi-transcendental structure *can* be identified in Kant's text, just as plainly as we could hope to see it. For the four headings correspond directly to the three parts of Walker's formulation of the typical structure of a transcendental argument, with the two premises reversed and the first duplicated:

> *Section I*: "Concerning the Original Predisposition to Good in Human Nature" (Premise 1a: An *original predisposition to good in human nature* is a necessary condition for the possibility of religious experience[78])
>
> *Section II*; "Concerning the Propensity to Evil in Human Nature" (Premise 1b: A *propensity to evil in human nature* is a necessary condition for the possibility of religious experience)
>
> *Section III*: "The Human Being is by Nature Evil" (Premise 2: We have religious experience – manifested here as "sin" or *failed moral obligation*)
>
> *Section IV*: "Concerning the Origin of Evil in Human Nature" (Conclusion: Therefore, we know a priori that the *origin* of evil is not empirical, but rational, derived from the presence in human nature of a good predisposition and a corresponding evil propensity)

The above parenthetical explanations of how Kant's four section headings can be regarded as the basic components of a proof with the same structure as a typical transcendental argument do not establish once and for all that Kant actually *presented* these steps in the main text of the corresponding sections; and even if he did, this fact alone does not guarantee that his proof is successful. The *success* of the proof – an issue I shall not attempt to assess in this essay – will depend on whether Kant succeeds in establishing as

true the two claims advanced in Sections I and II and on whether the reader has a mind to challenge the general claim that human beings are evil, as advanced in Section III. If he is successful in both respects, then the logical form of a transcendental argument will justify the conclusion here attributed to Section IV, that evil is *rooted* in the very structure of human reason. Instead of attempting this tall task here, my remaining goal will be merely to scan the content of Sections II and III for evidence that those sections really do advance the two premises of the "missing proof" I have been discussing in this essay.

Kant begins Section II by making a distinction between three aspects of the propensity to evil that directly parallels his distinction in Section I between the three aspects of the predisposition to good. I shall not discuss the implications of this parallelism here, except to say it supports the notion that Kant regarded these two sections as defending two sides of the same "coin" of human nature, just as we would expect if their arguments were intended to defend two aspects of the same premise in the overall proof being advanced in Part One. Significantly, the paragraph immediately following his account of the third level of evil (the "depravity … of the human heart") makes a direct reference to the *proof* being developed (*Religion*, 30):

> It will be noted that the propensity to evil is here established (as regards actions) in the human being, even the best; and so it also must be if it is to be proved that the propensity to evil among human beings is universal, or, which here amounts to the same thing, that it is woven into human nature.

Kant obviously thought his proof should render the evil propensity both necessary ("must be") and "universal". These are the very features to be "established" by any successful transcendental argument.[79] Kant's metaphor of *weaving* therefore seems to imply that he regarded this evil propensity as a synthetic a priori component of what it means to be human. As such, we may reasonably conclude that Kant believed his argument in Section II fulfilled the requirement of Premise Ib in the overall formal proof he presents in Part One,[80] as suggested above.

Section III is where Kant makes the highly misleading statement quoted at the outset of this essay. If my argument so far is correct, then he did not mean to imply that he *really would* "spare" himself the trouble of constructing a "formal proof". He meant only that the reality of human evil is so deeply

imbedded in human experience that it would be *possible* to (we "can") forgo such a philosophical formality. After reciting his various empirical examples of evil deeds, Kant reminds us that the goal of this step in his overall proof is strictly limited (*Religion*, 35):

> But even though the existence of this propensity to evil in human nature can be established through experiential demonstrations of the actual resistance in time of the human power of choice against the law, these demonstrations still do not teach us the real nature of that propensity or the ground of this resistance; that nature rather … must be cognized *a priori* from the concept of evil … What follows is the development of this concept.

The remainder of Section III basically rehashes what was already stated in Section II, only in light of the newly presented *empirical evidence* for treating human evil as *real*. The "experiential demonstrations" in the first half of Section III therefore perform the function of Premise 2 in the quasi-transcendental argument whose basic form I outlined earlier in this section. Kant's "development" of the "concept" in the remainder of his Section III attempts to show how the two premises are linked together.

Of the various interpretations examined above (§§1-2), Allison's comes closest to recognizing the quasi-transcendental nature of Kant's argument in Part One. He writes: "Kant's operative assumption is that without such a ground [i.e., the evil propensity in human nature], … moral evil would be impossible."[81] Although this way of reasoning exactly follows that of the typical transcendental argument in the first *Critique*, Allison shows no awareness of this similarity. Instead of portraying it as one of the two *premises* of the "formal" (quasi-transcendental) proof Kant presents in Part One, Allison treats Kant's argument as a mere *assumption*. Kant is not *assuming* human nature is grounded in an evil propensity; he is *arguing* (in typical transcendental fashion) that *a general type of human experience* (i.e., experience of the sort that makes us religious beings [see note 71]) would be impossible if no such ground existed in human nature. The type of general experience at issue here is *humanly perpetrated* (i.e., freely chosen) *evil*. The fact that we observe such evil in our perusal of the empirical world does not merely justify an assumption; it *constrains us to recognize* that human nature has been infected by evil at its very root.

I shall now conclude by addressing just a few of the many possible implications that arise out of this way of understanding Kant's proof of the evil propensity in human nature. First, if Kant really does

advance a quasi-transcendental argument in Part One of *Religion*, then would this mean he has compromised the freedom of choice that makes us responsible for the evil deeds we perform? Would a necessary and universal propensity to evil indicate that *we cannot help* but commit evil deeds and that we therefore should not be blamed? Whatever else he may wish to establish in Part One, Kant clearly and repeatedly affirms that his theory *must* be one that preserves this element of freedom. As Morgan rightly explains, "the fact that we did not bring the propensity to evil upon ourselves through a choice we could have avoided does not mean that it is not our free doing, and so it is still something for which we can be appropriately condemned, as the epithet 'evil' clearly indicates."[82] But this assertion does nothing to *explain* how Kant thinks he can get away with such an apparent sleight of hand.

Allison recognizes more openly that Kant's theory of the propensity to evil (even in his reconstructed versions) poses a definite threat to the freedom of the human moral agent, thus risking the self-destruction of Kant's entire moral system (along the lines bemoaned by Michalson and celebrated by Fenves). He attempts to diffuse this threat by pointing out that for Kant freedom is essentially a "causality of reason" and therefore does not always involve the capacity to do otherwise.[83] However, this threat can be abated even more persuasively by acknowledging the quasi-transcendental status of the proof Kant constructs in Part One. If the evil propensity is *transcendental*, not in any sense empirical – if it is not a real choice made at a specific point in a person's moral development, but a "choice" that can be *seen to be necessary and universal* wherever and whenever *any* evil act (even, Wood's protests notwithstanding, a *single* evil act) appears – then freedom is preserved. For the claim that the free choices we make in our everyday (spatio-temporal) experience are related *transcendentally* to an underlying "law of choice" that each agent *must have freely chosen* does not strip the moral agent of his or her freedom any more than the claim that causal laws in the phenomenal world are all governed transcendentally by the law of causality robs specific empirical causes of their thoroughly contingent nature.[84]

This way of understanding Kant's basic argument in Part One, as an attempt to establish a quasi-transcendental conclusion about the synthetic a priori boundaries that define the possibility of being religious, can also shed new light on Kant's notorious theory of "timeless choice". Kant seeks to preserve the freedom of the religious person (the person to whom the arguments in *Religion* are primarily

addressed) by claiming that, although the propensity to evil is "woven" into the very fabric of human nature, each human person nevertheless *chooses* to adopt the self-interest conveyed by this propensity into the fundamental maxim governing his or her disposition, thus incorporating evil into the very ground of his or her character. Kant's references to such a "timeless choice" simply mean that (as hinted in *Religion*, 20), when we observe ourselves (or any moral agent) performing even a *single* evil deed in the phenomenal world, we are justified in inferring from that deed that our character has been influenced by an evil propensity. Saying this "act" is "timeless" merely means, as Allison points out, that the "propensity cannot be thought as self-consciously adopted at a particular point in time. On the contrary, it is found to be already at work when moral deliberation begins and must be presupposed in order to conceive the possibility of immoral actions in beings for whom the moral law provides an incentive."[85]

A similar approach to proving the transcendental necessity of an evil propensity in human nature would be to regard it as the necessary *first free act* any moral agent can perform. Kant himself hints at this possibility at several points in *Religion*, as when he emphasizes the serious fault in Stoic ethics[86] by pointing out that, even from the time of our very first conscious choice, we (children and adults alike) are aware of a deep lack of moral integrity, that we are not beings who answer to a single moral principle but beings *stuck in a struggle* (*Religion*, 58n):

> For no matter how far back we direct our attention to our moral state, we find that this state is no longer *res integra*, and that we must rather start by dislodging from its possession the evil which has already taken up position there (as it could not have done, however, if it had not been incorporated by us into our maxims). That is, the first really good thing that a human being can do is to extricate himself from an evil which is to be sought not in his inclinations but in his perverted maxims, and hence in freedom itself.

Kant here claims that our first awareness of making a genuinely good moral choice already assumes the presence of evil, an evil propensity we have freely chosen to obey. If this is the case, then obviously *every person's first conscious moral choice is evil*.[87] In the earliest stage of personal development a child (or child-like person) might make technically good choices (choices that happen to be consistent with the moral law) merely because he or she is not aware of any inward potential for transgressing the moral law;

yet such "good" acts are of no moral worth, for they are grounded in mere assent to an unopposed (and therefore unconscious) demand. By contrast, once a person has committed an evil act and has thereby awakened consciousness of the moral struggle, an opportunity for genuine virtue (doing good in the face of an evil alternative) becomes possible.

We can now also see more clearly why anthropological research is so important to Kant's claim that *all* human beings start out their moral development with a propensity to evil. Just as Kant suggests in *Religion*, 20, he believes he can prove that human nature as such (and therefore *every* human being, at least in principle) is exposed to the evil propensity by constructing a quasi-transcendental argument based on *any* given example of an evil deed, however banal (or depraved) it may be. What he admits he cannot prove is that there are *absolutely no exceptions* to this transcendental principle, any more than the transcendental necessity of the law of causality in Kant's theoretical philosophy guarantees a total absence of any uncaused events.[88] Because the possibility of an exception is always present in both cases, the Critical philosopher cannot rule out, a priori, the possibility of miracles,[89] though we must affirm – and this is the brunt of Kant's main argument in *Religion*, where he is trying to persuade philosophically skeptical theologians and religious leaders to *give up* irrational beliefs and practices – that *such exceptions can play no constitutive part in our understanding of the rational foundations of human experience*, whether it be scientific or religious in its orientation. Despite Michalson's protests to the contrary, such apparent "wobbling" on Kant's part does not imply "instability"; rather, as Wood rightly argues (though without admitting Kant's need for a transcendental proof), such subtleties attest to the deep significance his position has for our empirical study of the ethical behavior and character of human beings.

The foregoing identification of the basic structure of Kant's quasi-transcendental argument for the human propensity to evil does not by any means end the discussion of this important topic. On the contrary, by not even addressing the question of the *validity* of Kant's argument, I have opened up for interpreters a whole new set of problems to consider. Is this argument, as I have now identified it, a real transcendental argument? How, if at all, do "quasi-"transcendental arguments differ from standard (theoretical) arguments exhibiting virtually the same form? Does Kant succeed in establishing each of the steps in this argument in the text of Part One (or elsewhere in his writings)? Or are reconstructions such

as Allison and Morgan present still needed to fill gaps in Kant's exposition? Even if Kant's proof does succeed, can we reconstruct it in a more concise, less perplexing form (perhaps along the lines sketched above, at note 87) that would establish even more persuasively the conclusion Kant sought to defend, that a rational account of human free will *requires* us to portray human nature as inevitably burdened with a propensity to evil? And perhaps most importantly for theologians and religious believers, does anthropological research confirm that there are *no exceptions* to the general rule that people succumb to this propensity by actually *becoming* radically evil, or would it be possible for a real historical person to persevere, unswerving, in conscious, lifelong service to humanity's good predisposition?

References

Allison, Henry, "On the Very Idea of a Propensity to Evil," *Journal of Value Inquiry* 36 (2002): 337-348.

————, "Ethics, Evil, and Anthropology in Kant: Remarks on Allen Wood's *Kant's Ethical Thought*," *Ethics* 111 (2001): 594-613.

————, "Reflections on the Banality of (Radical) Evil: A Kantian Analysis," in Henry Allison, *Idealism and Freedom: Essays on Kant's Theoretical and Practical Philosophy* (New York: Cambridge, 1996): 169-182(??). Reprinted from *Graduate Faculty Philosophy Journal* 18.2 (1995): 141-158

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

Anderson-Gold, Sharon, *Unnecessary Evil: History and Moral Progress in the Philosophy of Immanuel Kant* (Albany: State University of New York Press, 2001).

————, "God and Community: An Inquiry into the Religious Implications of the Highest Good," in Philip Rossi and Michael Wreen (eds.), *Kant's Philosophy of Religion Reconsidered* (Bloomington: Indiana University Press, 1991): 113-131.

————, "Kant's Rejection of Devilishness: The Limits of Human Volition," *Idealistic Studies* 14 (1984): 35-48.

Bernstein, Richard J., "Radical Evil: Kant at War with Himself," in Richard J. Bernstein, *Radical Evil: A philosophical investigation* (Cambridge: Polity Press, 2002), 11-45. Reprinted from M.P. Lara (ed.), *Rethinking Evil: Contemporary Perspectives* (Berkeley: University of California Press, 2001): 55-85.

Burns, R.M., "The Origins of Human Evil," *Scottish Journal of Theology* 53.3 (2000): 292-315.

Caswell, Matthew, "The Value of Humanity and Kant's Conception of Evil", *Journal of the History of Philosophy* 44.4 (October 2006): 635-663.

Copjec, Joan, *Imagine There's No Woman: Ethics and Sublimation* (Cambridge: MIT Press, 2002).

Cramer, Konrad, "Kants Bestimmung des Verhältnisses von Transzendentalphilosophie und

Moralphilosophie in den Einleitungen in die »Kritik der reinen Vernunft«", in Hans Friedrich Fulda and Jürgen Stolzenberg (eds.), *Architektonik und System in der Philosophie Kants* (Hamburg: Felix Meiner Verlag, 2001), 273-286.

Fendt, Gene, "Innate Corruption and the Space of Finite Freedom", *American Catholic Philosophical Review* LXVIII.2 (1994): 179-201.

Peter Fenves, *Late Kant: Towards another law of the earth* (New York: Routledge, 2003), Chapter 4 ("Out of the Blue: 'On the Radical Evil in Human Nature'"), 75-91.

Firestone, Chris L., and Nathan Jacobs, *In Defense of Kant's* Religion (Bloomington: Indiana University Press, forthcoming 2008).

Förster, Eckart (ed.), *Kant's Transcendental Deductions: The Three* Critiques *and the* Opus Postumum (Stanford: Stanford University Press, 1989).

Kant, Immanuel, *Religion within the Boundaries of Mere Reason*, in Immanuel Kant, *Religion and Rational Theology*, ed. Allen W. Wood and George di Giovanni (Cambridge: Cambridge University Press, 1996), 55-215.

Lilla, Mark, "Kant's Theological-Political Revolution", *The Review of Metaphysics* 52.2 (December 1998): 397-434.

Michalson, Gordon, *Fallen Freedom: Kant on radical evil and moral regeneration* (Cambridge: Cambridge University Press, 1990).

Morgan, Seiriol, "The Missing Formal Proof of Humanity's Radical Evil in Kant's Religion**"**, *Philosophical Review* 114 (2005): 63-114.

O'Connor, Daniel, "Good and Evil Disposition", *Kant-Studien* 76 (1985): 288-302.

Palmquist, Stephen, "Quantum Causality and Kantian Quarks – A Defense of the Compatibility between Transcendental Idealism and Quantum Physics" (forthcoming).

————, "Philosophers in the Public Square: A Religious Resolution of Kant's *Conflict*", in Chris L. Firestone and Stephen R. Palmquist (eds.), *Kant and the New Philosophy of Religion* (Bloomington: Indiana University Press, 2006), 230-254.

————, *Kant's Critical Religion: Volume Two of Kant's System of Perspectives* (Aldershot: Ashgate, 2000).

————, *Kant's System of Perspectives: An architectonic interpretation of the Critical philosophy*

(Lanham: University Press of America, 1993).

Strawson, Peter F., *The Bounds of Sense* (London: Methuen & Co. Ltd., 1966).

Stroud, Barry, "Transcendental Arguments," in Ralph C.S. Walker (ed.), *Kant On Pure Reason* (Oxford: Oxford University Press, 1982): 117-131. Reprinted from *The Journal of Philosophy* 65 (1968): 241-56.

Tillich, Paul, *Mystik und Schuldbewusstsein in Schellings philosophischer Entwicklung*, 1912. Tr. Victor Nuovo as *Mysticism and Guilt-Consciousness in Schelling's Philosophical Development* (London: Associated University Press, 1974).

Walker, Ralph C.S., *Kant* (London: Routledge and Kegan Paul Ltd, 1978).

Wood, Allen, "Religion, Ethical Community and the Struggle Against Evil", *Faith and Philosophy* 17.4 (October 2000): 498-511.

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

————, *Kant's Moral Religion* (Ithaca, N.Y.: Cornell University Press, 1970).

Notes

[1] Quotes from this book are taken from the Cambridge Edition of Kant's works, translated by George di Giovanni as *Religion within the Boundaries of Mere Reason* in Immanuel Kant, *Religion and Rational Theology*, ed. Allen W. Wood and George di Giovanni (Cambridge: Cambridge University Press, 1996). References to this book will be included in the text as "*Religion*", followed by the pagination from volume 6 of the Berlin Academy edition of Kant's works. For a defense of my alternative translation of the title, see §VI.1 of Palmquist (2000).

[2] See e.g., Allison (1990), 154. Allison thinks Kant is saying "that the necessity for such a [formal] proof is obviated" by the provision of such evidence. For a typical example of the many commentators who cite this passage in passing and merely assume it means Kant is shirking his philosophical responsibility to provide a proper proof for his theory that human nature has a built-in propensity to evil, see R.M. Burns (2000), 294, 297, who says Kant defends his theory by "providing only travellers' tales plus a few anecdotes about the treachery of friends as supporting evidence." Bernstein likewise finds Kant's "analysis of evil" to be "disappointing" (36). Shortly before quoting the passage from *Religion*, 32-33, he writes (34): "If there is one lesson that we should have learned from the Critical Philosophy, it is that genuinely synthetic universal claims can never be justified by appeal to experience; their justification requires a 'deduction' – a proof. Yet, when Kant reaches this crucial stage in his exposition, when we expect some sort of proof or justification of radical evil as a *universal* characteristic of human beings, *no* such proof is forthcoming." After presenting the passage in full, Bernstein again laments (34): "Kant never gives – or even attempts to give – a *proof* of his controversial and bold claim that man is evil by nature." Michalson concurs (67; see also 46) that "there is utterly no way that Kant, above all, could legitimately generate a claim about an intrinsic feature of human nature from even the lengthiest list of empirical examples." And Fendt agrees that "Kant's 'long melancholy litany' of the past sins of mankind is … a rhetorically inappropriate way for him to argue" (192).

[3] Kant begins Section II by defining a propensity (*Religion* 29) as "the subjective ground of the possibility of an inclination …, insofar as this possibility is contingent for humanity in general." The "possibility" of this inclination has to be "contingent", otherwise the propensity would be a *predisposition* (something we cannot help but have and do not freely choose). But its "subjective ground" can (and, Kant argues, in the case of evil, *is*) necessary. Just as Kant calls our predisposition to good "original" because it "belong[s] with necessity to the possibility of [the human] being" (28), so also he calls our evil propensity "natural" in order to indicate "that this propensity belongs to the human being universally (and hence to the character of the species)" (29). After describing "three different grades of

this natural propensity" (29), Kant states that, "if it is to be proved that the propensity to evil among human beings is universal", then it "must be" "established (as regards actions) in the human being, even the best" (30). Later, in Section III, he argues that, just as reason cannot "extirpate within itself the dignity of the law itself" (35), so also because "there is in the human being a natural propensity to evil…, it is also not to be *extirpated* through human forces" (37). By adding that nevertheless "it must be possible to *overcome* this evil" (37), Kant is hinting what will be fully elaborated in Part Two, "that some supernatural cooperation is also needed to [a human being's] becoming good or better" (44). The necessity and universality of the evil propensity in human nature are in this way intimately bound up for Kant with the necessity and universality of religion itself – a point I shall further develop in §3.

Bernstein, 31, reflects the sentiments of many readers when he finds it "difficult to understand … the very idea of a 'propensity to moral evil.' It is extraordinarily paradoxical (if not incoherent) to claim that there is a propensity to moral evil that is universal … Yet this is precisely what Kant does maintain." Bernstein thinks that if it is both universal and freely chosen (32), "then we would have to say that the human species *qua* species freely chooses this propensity. It is not clear that such a thesis is even intelligible." Far from denying that human nature is paradoxical, Kant repeatedly affirms it, using words such as "inscrutable" to describe the thesis he is maintaining. The point of Kant's proof, therefore, cannot be to *resolve the paradox*, but rather to explain why we are *constrained* to embrace it. And for this, as I shall argue below, something like a transcendental argument is needed.

[4] Wood (1999), 286-289, discusses the issue of Kant's proof (or lack thereof) for the evil propensity. As Kant states in *Religion*, 36, "the difference, whether the human being is good or evil, must not lie in the difference between the incentives that he incorporates into his maxim … but in … *which of the two he makes the condition of the other*. It follows that the human being (even the best) is evil only because he reverses the moral order of his incentives in incorporating them into his maxims."

[5] *Religion*, 25. Wood (1999), 287, quotes this passage and claims Kant's later citation of the "multitude of woeful examples" *constitutes* the "anthropological research" Kant had in mind. But I shall argue in §4 that Kant's strategy is considerably different from the rather awkward (and hardly persuasive!) approach Wood attributes to him.

[6] Wood (1999), 286.

[7] Wood (1999), 287. Wood's early interpretation of Kant's position here can be found in Wood (1970), 219-226.

[8] Wood (1999), 287. See Anderson-Gold (1991).

[9] Wood (1999), 289, clearly (and properly) distinguishes between the social manifestation of evil in human life (as a matter of anthropological research) and the individual responsibility each person has for any evil choices he or she

makes. To this extent, he does not conflate the topics of Parts One and Three of *Religion*. My point here is merely that the question of how Kant *proves* the evil propensity belongs to Part One, not Part Three; yet Wood thinks Kant's only (viable) answer comes in Part Three. Kant's very brief reference to the social manifestation of radical evil in Part One (*Religion*, 27) hardly amounts to a *grounding* of his proof in that anthropological fact. My goal in this essay is not to deny the relevance of the social (anthropological) side of evil – for Wood is right that this plays a very significant role in Kant's thinking – but rather to show that the "proof" Kant repeatedly refers to in Part One has a wholly different structure and orientation.

[10] Michalson (1990), 28, 128; see also 1-4, 10, 61, 84-85, 89, 142 and *passim*. Michalson's overall strategy in interpreting Kant over the years has been to view anything he cannot understand in Kant's text as a reflection of inconsistencies that Kant allowed into his theory by wanting to defend necessarily conflicting value systems. He claims Kant "wobbles" between his conflicting commitments throughout *Religion* and thereby fails to present a single, consistent position on most of the issues he addresses. As a result, even though Michalson recognizes that Kant's argument for evil *ought* to be transcendental (31, 41, 46), he interprets Kant's alleged failure to present such a proof as but one example of Kant's "series of delicate balancing acts" (47) where it seems "unclear where balancing shades off into self-contradiction." Thus, he calls attention to the alleged "peculiarity of [Kant's] line of argumentation" (64) in "using the claim that the source of moral evil is freedom as a premise in generating the further claim that the source of moral evil is unknowable." Michalson claims that Kant's "conclusion … appears to deny his ability to know his own major premise." This results in "a frustrating conceptual logjam" (67), whereby "it is never clear why Kant thinks radical evil is universal, or the propensity to evil innate." Michalson confesses that his own approach merely "adds to the complexity" (129) of trying to understand Kant. For a thoroughgoing refutation of Michalson's dismissive approach to Kant, see Firestone and Jacobs. Of course, the whole situation would look much different if, as I shall argue in §4, Kant does, in fact, present an argument for these claims.

[11] Allison (1990), Chapter 8 ("Radical Evil"), 146-161.

[12] Allison (1990), 146.

[13] On the transcendental status of any synthetic a priori claim, see Palmquist (1993), §IV.2-3. On the range of possible relationships between transcendental claims and the special form of proof Kant calls a "deduction", see the various essays in Förster. Caswell, 637, is a typical example of a commentator who (following Allison) refers to Kant's argument in Part One as an attempted "deduction of radical evil." Bernstein, 240-241, by contrast, is typical of those who "fail to find Allison's reasoning persuasive" when he claims Kant requires a deduction to establish his

claims in Part One. As Bernstein points out (240): "There is not the slightest indication that Kant himself ever thought that such a deduction was necessary or even possible." See also note 14, below.

[14] The only "deduction" Kant explicitly presents in *Religion* comes just where we would expect to find it, if we attend to the architectonic structure of Kant's argument in the first *Critique* as compared to that in *Religion*: in Part Two, where Kant needs to establish the "objective reality" of the "prototype of perfect humanity" – i.e., what Christian theology calls the divine "logos". See *Religion*, 76. Such parallels with the argumentative structure of the *Critiques* indicate that Kant was actually attending quite carefully to his Critical concerns as he wrote *Religion*.

[15] Allison (1990) concludes §I of Chapter 8 by stating (152): "the conceptual apparatus for articulating the doctrine of radical evil was already in place in 1785." The new contribution of *Religion* was "to explain how the attribution of a propensity to evil is compatible with freedom (no small task) and to argue that this propensity is universal."

[16] Allison (1990), 152

[17] Allison (1990), 153.

[18] Allison (1990), 153.

[19] Allison (1990), 153. He adds: "It is rather that one finds that this is how one has been behaving all along." If Kant were to rewrite Part One from the vantage point of the twentieth century, I believe he might refer to this "intelligible" act as "unconscious", in something like a Jungian sense of the term, whereby we are still held accountable for who we are unconsciously, even though we do not *consciously* choose to be that way. To say this act of giving in to the evil propensity in human nature is "timeless" (as Kant does) is to say that (like the Jungian unconscious) it "is not to be viewed as performed at some specific point in one's moral development" (154), for it "is already at work when moral reflection begins."

[20] Allison (1990), 154.

[21] Allison (1990), 154.

[22] Allison (1990), 154.

[23] Allison (1990), 155.

[24] Allison (1990), 155. Cf. Michalson, 45. See also note 59, below.

[25] Allison (1990), 155.

[26] Allison (1990), 156.

[27] Allison (1990), 156.

[28] Allison (1990), 157.

[29] Allison (1990), 157.

[30] Morgan, 65.

[31] Morgan, 65.

[32] Morgan, 69.

[33] Morgan, 69. Morgan goes on to argue (70): "Yet if the propensity to evil is inextirpable, and the possession of such amounts to possession of an evil disposition, in such a circumstance a human being would possess both a good and an evil disposition. But this would be syncretist latitudinarianism, in flat contradiction to the rigorism Kant insists upon." Unfortunately, Morgan here conflates Kant's clear distinction between the propensity (the *universal tendency* of human beings to be evil) and the disposition (the *actual character* of a given moral agent). Bernstein, 26, while properly distinguishing between these two aspects of Kant's theory, complains that Kant "never explains why the disposition (*Gesinnung*) of human beings can be good or evil, whereas there is a propensity (*Hang*) *only* to evil." My interpretation of Kant's proof for the propensity (see §4) will clear up this problem, showing that Kant does, in fact, explain his rationale. In a nutshell, he argues that human nature *must* be so constituted that our predisposition is good and our propensity is evil, otherwise we would not have a *free choice* to adopt either a good or an evil disposition!

[34] When Morgan eventually explains which aspect of Kant's theory he must reject in order to uphold his reconstructed proof, he rejects a position I do not believe Kant himself ever defends (95): namely, "the claim that the propensity to evil amounts to the adoption of an evil disposition." After an elaborate explanation of why he thinks this claim (as opposed to Kant's rigorism) must be abandoned, Morgan concludes (100): "there is no solid argument for holding that we all must have embraced an evil fundamental maxim; hence Kant's insistence that propensity and *Gesinnung* [disposition] must be one and the same is groundless." But as I understand Kant's position in Part One, the propensity is the *inescapable tendency* we all have to adopt an evil disposition; Kant never explicitly identifies the propensity with the disposition, otherwise he would indeed have to contradict himself in Part Two when he claims that an evil-hearted person can become good. Morgan claims that Kant "insist[s] on what looks like a conflation" of the propensity and the disposition (98), but never succeeds in demonstrating that this conflation is anything other than his own interpretive misreading of the text. Ironically, by giving up this alleged "conflation", Morgan inadvertently gives up the very claim Kant's transcendental argument is supposed to prove: *that the evil propensity in human nature mires us* all (as a species) *in radical evil*.

[35] Morgan, 79.

[36] Morgan, 82.

[37] Morgan, 85. Morgan describes this view in a way reminiscent of Genesis 3: "What the will really yearns for is the kind of freedom and power possessed by a very different kind of will, the infinite unlimited will of God." Later (88), he quotes a suggestive passage from Kant's *Anthropology* that provides "an even clearer indication that Kant holds the human will as such to be subject to an incentive to limitless self-assertion." Morgan's account shows how Kant's view of evil is consistent with the biblical notion that human beings want to be like God (i.e., lacking in all external constraints); yet he never explains how this way of conceiving of evil amounts to a *transcendental* argument.

[38] Morgan, 86. Indeed, Morgan goes on to claim that "license and self-love turn out to be identical."

[39] Morgan, 86-87.

[40] Morgan, 87.

[41] For example, Morgan, 89, says he has *elucidated* the nature of evil without *explaining* its mystery (something Kant would insist cannot be done); he also provides a Kantian *psychology* of evil by describing "the primary motivation of the evil person, the self-assertive determination that no limits be placed upon the choices the agent may make. It preserves freedom and responsibility by locating that insistence in a fundamental choice of a free will." Even if we grant all of this to Morgan, none of it constitutes a transcendental argument for evil!

[42] Morgan, 87, refers to three different passages from *Religion* where Kant argues that the evil propensity must be necessary and universal, but none of these texts has any relation to Morgan's reconstructed proof. Ironically, in the first sentence following the paragraph where he refers to those passages, Morgan states: "Of course, it must be frankly admitted that the argument I have presented is simply not present in the text of the *Religion*." This is true of the genuinely reconstructed portion of Morgan's argument; but his reconstructed argument makes *no reference* to the crucial qualities of being necessary and universal, while Morgan's *only* references to these qualities come as direct references to Kant's own statements. Obviously, then, Morgan's reconstructed argument does not meet the requirements of the "appropriate proof" Kant believed he had presented – a proof whose basic structure I shall attempt to lay bare in §4. Morgan claims he has shown that "it makes perfect sense to say that the propensity is universal and yet imputable to each one of us individually" (94), and I agree he accomplishes that goal. However, the *clarification* he provides never goes beyond a mere analysis of what Kant's various theories imply. Morgan never even attempts to describe the formal structure of the proof itself, as I shall attempt to do in §4.

[43] Kant's evasiveness in presenting an explicit proof could help explain why he never actually uses the *word* "transcendental" anywhere in *Religion*. Perhaps he was aware (at some level) that using the most technical of all his

technical terms would cause readers familiar with his Critical writings to *expect* a formal proof that he was not sure he could (or had) provide(d). While this is a possible explanation, I shall recommend a different one in §3.

[44] Wood (1999), 287, 402. The most we can infer from the fact that a human will is unholy is that it *might* not "follow the moral law, not that it displays a *propensity not* to follow the law." Kant's further claims that this propensity is both universal and innate are therefore certainly not derivable from Allison's reconstructed argument. Wood thinks Allison (1996) adopts a better approach – though Allison's essay on Arendt and the banality of evil adds nothing new to the issue of how Kant *proves* the evil propensity. Allison (2001) and Allison (2002) respond to Wood's criticisms and update his own theory of the evil propensity, agreeing with Wood's emphasis on Kant's anthropology of evil while noting that Kant's theory of the evil propensity cannot be *reduced* to mere "unsociable sociability" (337). He charges Wood with the error of conflating what could be called "the pure and the schematized concept of a propensity to evil" (345-346). Yet even in this latter essay, Allison continues to regard Kant's reference to empirical examples of evil "as a rhetorical ploy" and his reference to a possible "formal proof" as merely a tantalizing invitation to construct what Kant leaves, at best, merely implicit (Allison [2002], 341). Never, as far as I am aware, does Allison acknowledge the fact that in *Religion*, 39n, *Kant claims he has provided the needed proof*.

[45] Morgan claims his "argument supplies just what is needed to complement Allison's…. My argument shows how the root of such wrongdoing is a deeply disturbing competitive standpoint taken by the agent towards the social world" (91). In this sense, Morgan's reconstruction is an intriguing synthesis of Allison's and Wood's approaches; yet it does not go any further than either of its predecessors in showing the basic structure of a transcendental argument *Kant himself* actually advanced in Part One. Indeed, Morgan's admission that his reconstruction relates to "the social world" suggests that it (like Wood's own approach) is more relevant to Part Three of *Religion*, where Kant defends the need for good-hearted persons to join together in a church, than to Part One (cf. note 9, above).

[46] Morgan, 111.

[47] Morgan, 100.

[48] Morgan, 100.

[49] Fenves, 75. The German "*bloßen*" can mean "mere", "naked" or (perhaps most accurately) "bare". See Palmquist (2000), §6.1, for further discussion of this issue.

[50] Fenves, 75-76.

[51] Fenves, 77.

[52] Fenves, 79. He adds: "human beings are corrupted at the root, regardless of how upstanding they may appear in

the light of day." Fenves is quite right to point out that Kant's theory of the propensity to evil in human nature attempts to account not only for outward or obviously evil actions, but also for the hidden evil of hypocrisy, whereby the "tree" looks good even though the "root" is diseased.

[53] Fenves, 81.

[54] Fenves, 82. Predictably, Fenves has a heyday with Kant's claim in *Religion*, 32-33, that he "'spares' himself the trouble of a 'formal proof'" in favor of "a series of material ones." For he points out that if we "can merely presuppose evil in every human being, then one is justified in mistrusting everyone – even the best, even 'I. Kant'" (82).

[55] Fenves, 83. In classic post-modern style, Fenves adds (84): "Distrusting 'On the Radical Evil in Human Nature,' in sum, is the condition, of trusting its thesis – and vice versa."

[56] Fenves, 84. That this emphasis on irony and paradox in Part One is not entirely a figment of Fenves' creative imagination, but has the power to convey real insight about the deeper implications of Kant's theory of morality and religion, becomes clear when he makes the ironic (but authentically Kantian) point that "only the conscientious are hounded by their own conscience", whereas "those who enjoy peace of conscience are not innocent but … precisely the ones who falsify their inner assertions" (85).

[57] Fenves, 89.

[58] Fenves, 88, calls Kant's claim to have provided a proof (*Religion*, 39n) a *lie*: or he opines, at least, that in making this claim Kant "hides. Or hides from himself the absence of any 'formal,' 'proper,' 'literal,' or 'authentic' proof of the judgment that separates out the human species, condemns it, and damns it, not to hell, but to hope." The odd twist in the final word of this quote accurately reflects Kant's ultimate purpose in arguing for the propensity of evil, for if an originally good human nature is not corrupted by evil, then no problem remains for religion to solve. While I also agree with Fenves that Kant appears to have *hidden* his proof, I do not believe this means it cannot be found; on the contrary, in §4 I shall attempt to uncover its precise location within the text of Part One.

[59] Allison and Morgan reconstruct their arguments on the assumption that Kant never even addresses the possibility that human beings might have a propensity to good (see note 24, above). But Kant does briefly mention this option, in the second sentence of Section II (*Religion*, 29), where he says a propensity "can … be thought of (if it is good) as *acquired*, or (if evil) as *brought* by the human being *upon* himself." Again, this "acquiring" seems to be Kant's way of hinting, even at this early stage in the book, that the assistance of an outside (divine) power may be required.

[60] For discussions on the full range of Kant's various uses of transcendental deductions, see the essays in Förster.

[61] For an account of the four perspectives that operate in each of Kant's philosophical systems (namely, the transcendental, logical, empirical, and hypothetical), see Palmquist (1993), §IV.3. For summaries of how Kant's deductions operate from the logical perspective in the first two *Critiques*, see §VII.II.B and §VIII.II.B.

[62] A selection of quotes confirming Kant's emphasis on the *possibility* of human nature in Part One of *Religion* is given in note 3, above. See Palmquist (2000), §VII.1 and §VII.2.A, for a further explanation and defense of the claim that the transcendental perspective always seeks to demonstrate the possibility of something empirical. In the first *Critique* the Transcendental Aesthetic adopts the transcendental perspective to establish that space and time are synthetic a priori conditions that must be assumed in order for any empirical knowledge to arise. In the second *Critique* the first chapter of the Analytic likewise adopts the transcendental perspective to establish that freedom is the synthetic a priori condition that must be assumed in order for any moral action to arise. In neither case does Kant attempt a deduction of the conditions assumed. Deductions, rather, are saved for the categories of the understanding and the categories of good and evil, coming in the second stage of Kant's argument in both *Critiques*. If *Religion* were to contain any deduction, therefore, we should look for it in Part Two, not Part One. And that is precisely where we find Kant's only reference to a deduction in *Religion* (see note 14, above). Exploring this topic, however, is beyond the scope of the present essay.

[63] Cramer, 285-286.

[64] See Strawson (1966) and Stroud (1968/1982).

[65] Walker, 14-27.

[66] Walker, 14. An important point to keep in mind is that this "experience" refers not to *specific* experiences, but to the general fact that experience is something we do have. The search for *p*, therefore, is the search for what *must* be true in order for "experience in general" to be *possible*. If Kant turns out to have a transcendental argument for evil, this could solve the complaint of Michalson (and countless others) that "Kant consistently mixes claims presented as known a priori and appeals to human experience" (149), provided the appeals to experience are meant to be *generalized* as to the *possibility* of our experiences of evil. O'Connor, 296f, takes Kant's special mixture of approaches to indicate that his arguments are neither empirical nor transcendental, not realizing that transcendental arguments *always* mix such features by the very nature of their form. Likewise, Copjec expresses "wonder at the readiness with which [Kant] accepts the fact of our wickedness" (138; see also 142), for although she recognizes that his central question is "how evil is possible given the fact of freedom" (139), she does not notice how closely this corresponds to the form of a standard transcendental argument.

Neglecting this possibility, Fendt sides with Michalson against Allison, and claims (191-192n): "neither an empirical argument (which gets generality, but not universality) nor a transcendental argument (which gets universality and necessity) is possible in support of the claim [that human nature is corrupt]." Fendt argues (192n) that Allison (and anyone who thinks Kant's argument is transcendental) must be wrong because to make evil necessary "would mean we are all evil, but that evil would not be moral, since it would be necessary." Yet this ignores the fact that Kant does say evil is necessary (see note 3, above), and that the transcendental argument refers only to the *possibility* of the experience of evil in general, not to any specific empirical deeds that individual moral agents actually choose. (Fendt himself quotes Kant's emphasis on "'the conditions for the possibility of' choosing evil" [192; see also 194], but fails to notice that this very phrase is used in Kant's standard transcendental arguments.) Fendt draws attention to a *paradox* in Kant's position that must be addressed by any viable solution. (Bernstein, 241, aptly summarizes the same paradox as amounting to "an absurd – indeed, self-contradictory – conclusion. All human beings (the human race or species) *necessarily* freely chose the propensity to moral evil." Absurd though it may seem, this is the view Kant defends.) To do away with the paradox entirely, as Fendt does by suggesting that Kant *should* have argued for the universality of evil merely by employing "[a]n *ad hominem* argument" to the effect that "he who is without sin should cast the first stone at the theory" (192), would surely not have satisfied a philosopher such as Kant, who always sought to support his key claims with rigorous proofs. Fendt thinks Kant is constantly tempted "to deal with sin as a metaphysical problem" (193n); but I shall argue in §4 that his aim is rather to counteract such tendencies among Christian theologians by showing how sin is a *transcendental* problem.

[67] Walker, 18-20. Walker contrasts his position with that of Wilkerson, who argues that the crucial step in a transcendental argument is synthetic. Examining the pros and cons of this debate is beyond the scope of the present essay, and I do not wish to take a stand on the matter here. However, a noteworthy point is that Allison explicitly acknowledges that his reconstruction of Kant's argument is analytic. If Walker is right, then Allison's reconstruction (or Kant's original, if it follows a similar approach) could easily be adapted to serve as *one of the steps* in a quasi-transcendental argument for evil. Thus, Fendt's claim that Kant "is giving a perspicuous analysis of the only ground on which moral evil can and does arise" and that this means he is *not* "transcendentally deducing the positing of evil" (194), ignores the fact that a transcendental argument might contain within it (as one of the premises) just such an analysis of the meaning of a concept!

[68] Walker, 21.

[69] Thus, for example, the various essays in Förster range throughout all three *Critiques* and the *Opus Postumum.*

[70] *Religion*, 10. The second edition Preface makes a similar point, explicitly appealing to the Christian tradition as the primary focus of Kant's religious examples (12-13), and reminding the reader of his interest in persuading the biblical theologian to think along philosophically respectable lines. He even implicitly affirms a critic's suggestion that the book's guiding question is (13): "How is the ecclesiastical system of dogmatics possible, in its concepts and doctrines, according to pure (theoretical and practical) reason?" – though he hastily adds that he will answer this (quasi-transcendental!) question in a way that does not require a prior acquaintance with the details of his Critical philosophy. I more thoroughly defend this way of portraying Kant's intentions, as oriented primarily toward theologians and pastors, in Palmquist (2006).

[71] A proper understanding and assessment of my argument in §4 requires the reader to understand (and accept!) that I am here using the phrase "religious experience" *not* in the way it came to be used in the wake of the post-Kantian Romantic movement and by philosophers of religion in the twentieth century – i.e., as an experiential basis for a new proof of God's existence. As used here, the phrase refers not to a direct (e.g., mystical) encounter between God and a human being (or humanity in general) – though I have argued elsewhere that Kant had leanings in this direction (see Palmquist [2000], Part Four). Rather, my use of "religious experience" here and throughout §4 refers merely to our awareness of empirically discernable beliefs and practices that have come to be meaningfully associated with a particular religious tradition. Despite the risk of being misinterpreted, I have preserved this usage (which, I admit, was *not* Kant's) in order to highlight the direct parallels between his quasi-transcendental argument in Part One of *Religion* and his typical transcendental arguments in the first *Critique*, where the word "experience" *is* used to refer to our awareness of empirically discernable beliefs and practices that come to be meaningfully associated with a particular *scientific* tradition. Lilla, 403, argues that "religious experience" was indeed Kant's focus in *Religion*, as a response to the "disturbing parallel" he had noticed "between Rousseau's moral sublimity and Swedenborg's ravings about the spirit world."

[72] Wood (1999), 286.

[73] Wood (1999), 402. It is important to note that Wood here reverses the order of the argument as Kant introduces it. Kant states not that we can infer a person's guilty *conduct* from a general (a priori) *principle*, but that we can infer the necessity of such a *principle* from any given example of *admittedly* guilty conduct. This clarification will turn out to be crucial to a proper understanding of the structure of Kant's quasi-transcendental argument. On Wood's reading of the quoted passage, any proof other than an empirical one would indeed be "wildly implausible".

[74] Wood (1999), 402.

[75] Wood (1999), 286. Wood's response to this concern is to emphasize Kant's appeal to anthropology (see note 5).

[76] Similarly, Allison (2002), 338, describes Kant as engaging "in a kind of thought experiment, the aim of which is to spell out just what we are committed to, if we take seriously the idea that evil is to be imputed…. More precisely, ["Kant's positing of an inscrutable propensity as the ground of evil"] marks the limit of such an analysis, the point beyond which there is nothing more to be said." Despite coming so close to identifying the precise nature of Kant's argument, Allison does not call attention to the fact that to create a "thought experiment" in this manner is to construct a quasi-transcendental argument.

[77] Palmquist (1993), Chapter III and *passim.* See also the chapter on Architectonic in the Doctrine of Method of the first *Critique*. A typical example of the tendency many commentators have to dismiss Kant's emphasis without even trying to understand what it involves is when Michalson (101) responds to one of the many claims Kant makes that he fails to understand by dismissing it as evidence that "Kant is having a sudden fit of architectonic nostalgia."

[78] For an explanation of the special meaning of this phrase, "religious experience", see note 71, above.

[79] For more examples of Kant's references to necessity and universality in Part One, see note 3, above.

[80] Morgan complains: "We never receive [in Part One of *Religion*, or anywhere else in Kant's works] a proper explanation of why the concept of a propensity to evil is supposed to be that of the determining ground of evil actions" (Morgan, 98). He further claims (100) that the fact "that the illicit [i.e., the evil propensity] exerts an inevitable pull on all human beings does not of course entail that everyone embraces evil as their fundamental commitment, merely that anyone may choose it, and we all feel an incentive to do so." If this is the case, however, then Kant's quasi-transcendental argument fails. For according to the argument I have advanced above, we could not be aware of this propensity in the first place, if we had not *actually* given in to it. That is, Kant's "explanation" of the evil propensity is *transcendental*: we experience in ourselves the "inevitable pull" toward evil; that pull is possible only on the condition that the universal propensity of human nature is toward evil; therefore human nature *must be* exposed to an evil propensity at its root. Bernstein, 33, makes a similarly instructive error when he writes: "Presumably, the introduction of the concept of radical evil is intended to explain *why* … we deviate from following the moral law. We do not always follow the moral law *because*, as human beings, we have an innate propensity to evil…. But does this 'because' really explain anything? Does it do any conceptual work? I do not think so." Like most commentators before him, Bernstein here presents Kant's argument in exactly the *reverse* of its correct form. Kant's logic is not "we are evil because our propensity is evil" but rather "the human propensity *must* be evil,

otherwise no human being *could* be evil."

[81] Allison, 147. However, in assessing Kant's references to empirical evil in Section III, Allison (154) claims "the *most* that this evidence can show is that evil is widespread, not that there is a universal propensity to it." This is not true if, as I have argued here, Kant intends his citations of these examples to constitute the second step in a quasi-transcendental argument. By coming so close to recognizing the location and nature of Kant's argument, but missing its most central point, Allison in a sense goes further astray than Wood. Wood's emphasis on the social nature of evil is entirely correct as far as it goes, *provided* he merely remains silent on the question of the *proof* in Part One, rather than claiming (as Allison and others impute to him) to *reduce* Kant's theory of evil to this social aspect. In other words, the position I am defending here is *less inconsistent* with Wood's position than with any of the others I have considered, so long as he merely claims ignorance of any a priori proof in Part One, rather than claiming it is both necessary and missing, as Allison and others do. Whereas my position, if correct, would *refute* the errors of Allison and the other interpreters mentioned in §2, Wood could respond simply by accepting my discovery as a potentially accurate reflection of Kant's intentions, then asking how this quasi-transcendental proof provides a rational grounding for his own emphasis on the social aspect of Kant's theory (see note 9, above).

[82] Morgan, 94.

[83] Allison (2002, 343) notes, for example, that according to Kant the fact that God cannot choose to disobey the moral law does not make God's choices unfree.

[84] Specific causal laws, for Kant, cannot be *derived (or deduced) from* the law of causality, even though they derive their necessity and universality from their transcendental dependence on that law. I am suggesting that Kant is advancing exactly the same *type* of argument with regard to the relationship between evil actions and the propensity to evil that he claims is "woven into human nature."

[85] Allison (2002), 341. Again, this emphasis on "possibility" shows that Kant's argument is quasi-transcendental – though Allison never acknowledges this.

[86] Kant thinks the Stoics treated inclinations too harshly, for they are not evil *in themselves*, and thus do not always need to be denied; what is evil is the disposition whereby a person actively prefers inclination.

[87] This idea, presented here in seed form, was later nurtured and developed into a full-fledged theory by Schelling. Twentieth-century theologian Paul Tillich wrote his doctoral dissertation (Tillich [1912]) on the experience arising out of the notion that human beings "fall" into consciousness, especially *guilt*-consciousness. That is, human consciousness itself *first arises* as a direct result of sin: our ability to know arises only because we become aware

that we have transgressed the moral law. Thus, Copjec is surely right to say (143) Kant "is arguing that our only consciousness of the law is our consciousness of our transgression of it." Or as Bernstein puts it (410): "*Homo religiousus* is *homo cogitans* in action." Whereas Kant himself only hints at such claims, they are surely quite consistent with (and help to flesh out) the arguments he defends more fully in Part One of *Religion.*

[88] See Palmquist (forthcoming) for an extended argument that the lack of causal interactions in quantum physics renders it incompatible with Kant's theoretical philosophy.

[89] *Religion*, 89n, where Kant says "miracles must be admitted as [occurring] *daily* … or else *never* …". For a discussion of this point, see Palmquist (2000), 474-477.