

*Forthcoming in:* Sharon Anderson-Gold and Pablo Muchnik (eds.), *Kant's Anatomy of Evil: Interpretative Essays and Contemporary Applications*. Cambridge: Cambridge University Press, 2009.

Word count 8515

## **Kant and the Intelligibility of Evil**

Allen Wood

Kant's reasons for inquiring into the radical evil in human nature are very different from those that might now lead us to ask questions about evil. The aim of *Religion within the Boundaries of Mere Reason* was to explain to an audience of Christians (of eighteenth century Lutherans) how their faith might be reconciled with a rational Enlightenment morality.<sup>1</sup> Radical evil is the book's point of departure because of the religious importance of the Christian doctrine of sin. In Part One of the *Religion*, Kant's aim is to articulate that doctrine in rationalistic terms, so as to show in the other three parts how the Christian doctrines of justification and atonement, as well as the function of the church and revelation, might be articulated within the framework of a moral philosophy based on the autonomy of reason.

Today such aims make Kant far more enemies than friends. Christians, and religious people generally, typically charge him with "watering down" the faith, or even with offering their religion a philosophical Trojan horse concealing within it the entire army of modern secular unbelief.<sup>2</sup> On the other side, unsympathetic secular philosophers view the *Religion* as proof that Kantian ethics is at bottom nothing but traditional superstition. Both these reactions seem to me utterly wrongheaded, but here I will not address either of them directly. Instead, my purpose will be to see how Kant's reflections on evil might speak to concerns that are more likely to interest us.

The “evil” in question here not just the bad things that *happen* to people – the pain, grief and sorrow, injury, starvation, death, even their feelings violation and humiliation. For Kant, all these would all fall under the heading of “ill” (*Übel*), or human unhappiness. “Evil” (*Böse*) refers to something human beings *do*. More precisely, it consists in actions that they should not do but choose to do, and the principles that lead to these choices. ‘Evil’ includes acts of violence and cruelty -- war, rape, conquest, torture, terrorism, genocide – as well as lesser acts of cruelty, callousness, degradation, and disrespect for humanity – in fact, anything people do when they violate their duties and fail to live up to the dignity of their rational nature. ‘Evil’ also refers to our social practices. It includes the obscene gap between rich and poor, both within each society and between different societies, and the oppression of the powerless, based on these economic evils, on social customs, or the abuse of power built into political systems. Evil certainly includes as well what the human species inflicts on itself and other living things through its irresponsible relation to the natural environment.

When I speak about *our* questions concerning evil, what I mean is such questions as these: How can people do such things to one another, and even to themselves? What, at bottom, does such conduct consist in? And how, if at all, can we make sense of it? What meaning might it have for us? The answers to these questions are what I mean in the title of this paper by ‘the intelligibility of evil’.

**Can evil be made intelligible?** Even as we ask these questions, however, we also ask ourselves whether they make any sense. Perhaps our questions are nothing but a rhetorical expression of anger and despair, and taken literally, admit of no answers at all. According to one way of looking at the matter, “evil” is simply a word we use to express

attitudes of disapproval, blame, or horror at certain deeds. These deeds, and the choices of those who do them, are natural or social facts that have their physical, psychological or social explanations. For those who hold that rationality is only a matter of whether you select the right means to satisfy whatever aims or desires you happen to have, evil deeds may even be completely rational. Once we understand why these deeds occur, the only thing left to explain is why we take the negative attitudes toward them that we do: why we consider them “evil”. And there will also be psychological explanations for our attitudes too -- maybe rational explanations, maybe not. Put together the explanations for evil deeds and the explanations of our attitudes toward them and you have made evil intelligible in the only sense evil could be made intelligible.

This “deflationist” view of evil, however, remains totally unresponsive to our questions. It cuts the Gordian knot by making our questions about evil disappear, because in the most straightforward sense it makes *evil itself* disappear from the world. It tells us, in effect, that our questions about evil do not constitute any genuine inquiry at all – maybe, as already suggested, they are only rhetorical outbursts. Perhaps they too express simply an all-too-human attitude or mood that we must simply shrug at, recognizing it as part of our psychology. Or we may even accept the deflationist view of evil as having a certain sublimity, sharing something in common with a view of life we find in philosophers such as the Stoics and Spinoza, who take it to be the part of reason to rise above our emotional attitudes toward evil, to overcome them through what Spinoza called *amor intellectualis Dei*. Either way, with such a view our inquiry into evil would reach a dead end or be diverted into an inquiry about something else.

I won't try to refute such views directly on their own terms. But I will proceed on the assumption that they are wrong. For the apparent sublimity of the Stoic or Spinozist transcendence of our attitudes toward evil has exactly the same cause as the evident shallowness and untenability of the deflationist rejection of our questions about evil. None of these views, namely, are anything that we human beings could ever unite with our reflective experience of human life. They are views for gods, or perhaps for robots, bungled experiments at replicating humanity, or spectacularly successful attempts at constructing something simpler -- beings whose artificial intelligence has been truncated -- or, if you prefer, purified -- so that it entirely lacks those rational capacities that make us human beings, moral beings who care about our lives and those of others.

I will take for granted, therefore, that there really is evil, that we are right in asking why it occurs, and wrong to think that it is the part of reason to rise above this question or dismiss it as meaningless. Part of what it means to assume the reality of evil is to assume, along with Kant, that doing evil is *contrary to reason* – that is, that evil is something we have decisive reasons for not doing. If an action is what I have most reason to do, then there seems no longer to be any rational force in the assertion that I should not have done it – that it is really evil. If we don't assume this, then such an assertion seems only to express a certain irrational (or at least non-rational) negative attitude toward the action -- a reaction, moreover, that the agent has no reason to take seriously. So we are back with the view that there really are no evil actions, only a set of natural occurrences (in themselves neither good nor evil) and our non-rational attitudes toward them.<sup>3</sup>

Yet as soon as we grant the reality of evil, we immediately face some familiar and serious problems about how it could ever be explained or made intelligible. For evil is a

species of *rationaly motivated unreason*. Evil is like self-deception or *akrasia*, which notoriously give rise to paradoxes about how to understand them or perhaps even how to provide them with a coherent description.<sup>4</sup> For Kant, in fact, self-deception and *akrasia* fall under the heading of evil. About self-deception (or what he calls “the inner lie”) Kant acknowledges that there are difficulties understanding how it can be possible, but he has no doubt that it plays a large role in human life and also that it is a violation of a duty to oneself (MS 6:430-431). *Akrasia* belongs to that “frailty” of the will Kant cites the first or lowest degree of the radical evil in human nature (R 6:29). The paradoxes involved in *akrasia* and self-deception are therefore included, at least in part, in the more fundamental problem of the intelligibility of evil.

The basic problem about the intelligibility of evil can be stated in the following dilemma: There are apparently only two things we might mean by “explaining” evil or “making evil intelligible”. One would be an explanation of it as an action that is done for reasons. The other would be a causal explanation of it as arising from antecedent conditions. Either kind explanation, however, if fully successful, would thereby abolish what is evil about the action or display it as something that was not evil after all.

An evil action can be understood as done for reasons in a limited sense. For instance, it can be the action that is the best means to what I want most, or the action that will contribute most to my happiness. But by hypothesis, whatever reasons I might have for doing an evil action, there are moral reasons for me not to do it, and these are decisive reasons for me not to do it. (This is simply what it *means* for the action to be evil.) But then in principle there could never be a fully satisfactory explanation of an evil action as an action for reasons. An explanation that is not a rational explanation however -- a

causal explanation, for instance – would be incapable of making intelligible precisely what is *evil* in the action, because evil is conceived precisely in *rational* terms -- as a rationally motivated yet contra-rational action. Hence a causal explanation could never get at precisely what is *evil* in the action. Further, a causal explanation would apparently show how antecedent conditions made the action necessary, hence any other action impossible for the agent. But that would also do away with the agent's responsibility for the action -- and with it, the possibility of the action itself *as evil*.

These intractable difficulties are no doubt part the appeal of the deflationist view of evil as well as the perennial appeal of Socratic paradoxes about *akrasia* and the endless problems philosophers have in conceptualizing self-deception. But if someone were to argue on such grounds that self-deception is impossible, then the inevitable rejoinder -- utterly decisive -- would be to accuse this person of being self-deceived in that denial. And an analogous objection to deflationism about evil would be more powerful than any that could be brought against the attempt to make evil intelligible on the basis of the dilemma just presented. We have no choice, then, but to persevere in our assumption that evil is real and then see how far, and in what ways, such actions might still admit of being made intelligible. The point to appreciate going in is that no particular attempt to make evil intelligible should be dismissed simply because it runs afoul of the formidable difficulties just mentioned. For these difficulties simply come with the territory; they are not defects of some specific attempt, such as Kant's, to understand or explain evil. In fact, those who might expect evil to be made fully intelligible in either rational or causal terms don't even have a coherent conception of what they are asking for.

Kant exhibits a full awareness of these difficulties. He repeatedly emphasizes the limits in principle of both the attempt to conceive and to explain evil. He says the source of evil must lie in the free choice of the rational being, the choice to adopt an evil maxim. But he also insists that “there cannot be any further cognition of the subjective ground or the cause of this adoption (although we cannot avoid asking about it)” (R 6:25). “We are as just as incapable of assigning a further cause for why evil has corrupted the very highest maxim in us, though this is our own deed, as we are for a fundamental property that belongs to our nature” (R 6:32). Empirical evidences of the existence of a propensity to evil “do not teach us the real nature of that propensity or the ground of [our power of choice’s] resistance [to the moral law]” (R 6:35). Our choice of evil in time “cannot be derived from some *preceding* state or other” (R 6:39). We cannot even “inquire into the origin in time [of an evil deed], but must inquire only into its origin in reason” (R 6:41). That is, we cannot inquire into the *cause*, but only into the *character*, of a freely adopted maxim of evil choice.<sup>5</sup>

The first thing we need, then, is a coherent conception of what might count as making evil intelligible. There are, in fact, two things that might count as doing this. The first we might call ‘forming an intelligible concept of evil’. It consists in conceptualizing evil choices as following a highly general pattern that, although not fully rational, is nevertheless to a degree rational and also familiar to us as a way that human beings do in fact commonly choose. The Kantian name for this task is *identifying the fundamental maxim of evil* – or, for short, the *maxim problem*. Evil might be made still more intelligible if we could understand this general pattern of less than fully rational choice as fitting into human nature as it shows itself under the conditions in which human life has

developed on earth. This would help us to understand the persistence and prevalence of evil as a fact of human life, and also enable us to attach a meaning to evil, which might orient both our understanding of it and our struggle against it. This task is what Kant sets himself when he tries to identify evil *as a human propensity (Hang)*. So we can call it the *propensity problem*. Let us consider Kant's solution to these two problems in turn.

**The maxim problem.** Kant distinguishes three original "predispositions" (*Anlagen*) that belong to human nature: 1. animality, 2. humanity and 3. personality. None of these, he says, is inherently evil, and all may be regarded as present in us *for good* (R 6:28). *Animality* is the original source of our natural or instinctive impulses, hence of all our empirical desires or inclinations, including first, for "mechanical self-love" (self-preservation), second, for "propagation of the species" (the sexual drive) and third, for "community with other human beings (sociability) (R 6:26-27). *Humanity* is the rational capacity to set ends and devise means to them, and also the capacity for rational *self-love*, or the pursuit of our empirical ends as a whole, under the heading of happiness (R 6:27). This is the predisposition that first achieves development in society, through the *cultivation* by education of our skills to pursue ends, and then through the *civilization* of our nature through association with others, which shapes and modifies our conception of our well-being by comparing our state with that of others. Finally, *personality* is our capacity to respect the moral law, as the fundamental rational principle of the will, and to make that respect a sufficient incentive for obedience to the moral law (R 6:27-28). It too is a predisposition that is developed in the social condition, by the process that (parallel to those of cultivation and civilization) Kant calls 'moralization' – but this is a process, in his view, that human history has barely begun (VA 7:326-327)

None of the three predispositions is in itself evil, but evil must arise from a propensity we display in their use or exercise. Yet evil can also not be traced to either the first or the third predisposition. Our natural instincts (or animality) involves no principle of choice – which alone can be good or evil (R 6:34-35). They are in themselves innocent, and are capable of being involved in evil only insofar as we incorporate them as incentives into a freely chosen maxim (R 6:24). But then it is this choice, and not its instinctive source, that is good or evil. Evil also cannot be traced to our predisposition to personality -- our original relation to the moral law, as if we might have a basic incentive to disobey rather than to obey it – what Kant calls a “diabolical will” (R 6:35). Instead, the fundamental maxim of evil – Kant’s solution to what I have called the *maxim problem* – is that evil lies not in which incentives we incorporate into our maxim, but in the order of priority among them. As moral beings, we have rational incentives to action both in the moral law and also in our inclinations and self-love. Evil is conceivable only in the form of a maxim, or freely chosen subjective principle of the will, which inverts the rationally correct order of these incentives, and involves the preference of the incentives of inclination or self-love over those of morality. Moral goodness consists in acting on incentives of inclination and self-love but only on the condition that they agree with the moral law. Evil, on the contrary, consists in subordinating the incentive of morality to the incentives of inclination and self-love (R 6:36-39).

Kant’s view here, especially his rejection of the possibility of a “diabolical will,” is sometimes criticized for not allowing for the possibility – as it is put -- that people can do “evil for evil’s sake.” The objectors think that Kant is denying we can choose an action not because it promotes our self-interest or satisfies some contingent desire, but simply

*because it is wrong.*<sup>6</sup> But I think they misunderstand what he is claiming. Kant's position is that it would be incoherent to suppose that a being could be responsible for obeying the moral law and yet lack any rational incentive to obey the law, possessing originally *only* a rational incentive to *disobey* it (or, also that a being might originally have two directly contrary rational incentives, which would involve the supposition that the being's rational faculty itself is self-contradictory and incoherent). That is what Kant denies under the heading of a "diabolical will". The point is that there could be no ground to condemn the choices of such a being for going against morality, since it would have no reason, no capacity, to choose in favor of morality. Whatever harm to human or other beings might be caused by its actions, they could not be condemned as evil.<sup>7</sup>

When Kant denies that human beings can "incorporate evil as evil for an incentive into their maxim" (R 6:37), we easily misunderstand this if we assume a certain moral psychology, and a conception of moral reason, that is very different from his. Kant holds that for a rational being, the moral law simply as such is a rational incentive; no distinct (empirical) inclination (such as sympathy, or some desire for social conformity conditioned in us by our upbringing) is needed to give us an incentive to obey it. (This lies at the heart of Kant's thesis that the moral law is a law of *autonomy*, self-legislated by our own reason.) Kant's denial that "evil as evil" can be an incentive for us is the denial that anything parallel to this could be true in the case of evil – in other words, that we might have an original *rational* incentive to disobey the law -- or, as he also puts it, that we could have an "evil reason" (R 6:35). Unlike the original rational incentive to obey the moral law, he is claiming, our incentives to disobey it must take the form of empirical inclinations that provide an incentive to disobedience.<sup>8</sup>

Kant does not deny, however, that these inclinations can attract us to conduct that is directly contrary to what morality requires (that they might be *empirical desires* for “evil as evil”). For example, the moral law requires us to make the happiness of others our end, and so forbids us to take their unhappiness as an end for its own sake. What Kant calls the “vices of hatred” -- envy, ingratitude and malice – are vices because they involve making the unhappiness of another directly an end (MS 6:458-461). This looks like “evil for evil’s sake” if anything could be.

We all know that people can act “self-destructively” in the sense that they systematically do the very opposite of something they fundamentally will. For example, there are people who directly will to frustrate their own happiness – by becoming addicted to drink or drugs, or getting involved in abusive relationships with others. Likewise “doing evil for evil’s sake” could be considered a case of *moral* self-destructiveness, where someone chooses to disobey the moral law simply because they know that obeying it is what they ought to do. The choice so to behave would be based on an inclination to defy the moral law, and this inclination would be which the agent has given priority to the rational motive to obey the law. Nothing in Kant’s denial of a “diabolical will,” therefore, involves the denial of moral self-destructiveness. We can see that this is so once we realize about self-destructive patterns of motivation that the person always also (and more fundamentally) *wills* the thing their self-destructive behavior acts against. Those who act self-destructively in regard to their own happiness do it because they *also* (and more fundamentally) will to be happy. If they did not, they would not have the self-destructive motive to make themselves unhappy. And the morally self-destructive person, who does something “because it is wrong” likewise has a fundamental incentive

to do the right thing; otherwise, their action directly against morality would not have the meaning that it does. An act of malice toward another, for example, is malicious precisely because the agent knows that morality tells us to benefit others and not to harm them. Without the moral incentive in the picture, the act might still be harmful, but it would not count as a *malicious* act. Far from denying the possibility of “doing evil for evil’s sake”, Kant’s position is needed in order to give a correct account of what it is to do this.<sup>9</sup>

It is true that in the *Religion*’s discussion of evil Kant does not bother to distinguish between evil actions we might do because they benefit us at another’s expense and evil actions that we do precisely in order to harm the other (whether we benefit from them or not, or even are ourselves harmed by them). Here as elsewhere, he sometimes tends to emphasize the contrast between moral motivation and non-moral motivation, at the expense of various other contrasts between different species of non-moral motivation. His aim, after all, is to capture the most fundamental maxim of evil, which necessarily involves bringing non-moral motivation under a single heading (“self-love,” “inclination”, “the incentives of our sensuous nature”) (R 6:36-37). Perhaps this makes it easier for us to think that he is reducing all evil maxims to some one type – that of the self-indulgent hedonist, for instance, or the self-interested schemer – and that he is excluding others, such as the self-righteous hypocrite, or the malicious person consumed by spitefulness or hatred. But the aim in identifying the underlying maxim of evil is not to reduce all evildoers to a single human type, but instead only to conceptualize what is involved in acting against moral reason – in the common human pattern of volition of which the motivated unreason of evil consists. We miss the point of Kant’s account if we don’t recognize that it acknowledges that non-moral incentives take very different forms,

some shrewdly prudential, some vengeful and malicious, some involving disguise and self-deception, as when evil assumes the cloak of arrogant self-righteousness or religious hypocrisy.

**The propensity problem.** So far we have seen only Kant's solution to the *maxim problem*. The maxim of evil is to invert the rational order of incentives, placing self-love or inclination ahead of morality. This makes evil intelligible to a degree, because it is often consistent with both instrumental and prudential rationality. Moreover, it follows a pattern in human choice that is entirely intelligible in the sense that it is familiar to all of us, both in our own conduct and in the conduct of others. This provides us with an intelligible account of *what evil is*. What remains, however, is an even more difficult problem -- *the propensity problem*, the problem of understanding the prevalence of evil in the world, and its meaning in human life.

By the word 'propensity' (*Hang*) Kant means "the subjective ground of the possibility of an inclination (habitual desire, *concupiscentia*) insofar as this possibility is contingent for humanity in general" (R 6:29). A common example of a propensity is the propensity to consume intoxicants, which is aroused in some people by acquaintance with them (R 6:29n). A propensity is an empirical pattern of choice, or a desire to choose according to a determinate maxim. The propensity to evil is a propensity for the contra-rational choice that inverts the rational order of incentives, placing the incentives of self-love or inclination ahead of those of moral reason. This propensity is familiar enough to us empirically. In the attempt to understand evil, the *propensity problem* is that of coming to understand what it means that we have the propensity to evil, and why it is so prevalent among us human beings.

**The social origin of evil.** Kant's solution to the propensity problem is not highlighted in Part One of the *Religion*, because the aims of his discussion are those I have described, and not *our* aims in asking about evil. The propensity problem is also of only marginal interest to Kant in the Part Two, and begins to play a significant part in his aims only in Part Three of the *Religion*. Nevertheless, Kant's solution to the propensity problem is presented in the *Religion* both clearly and emphatically, and it coheres with his anthropology and philosophy of history as presented in other works. This solution is that *the human propensity to evil arises in the social condition, and develops along with the processes of cultivation and civilization that belong to it*. Though it is not emphasized in Part One, the social origin of evil is clearly indicated in Kant's remarks about the predisposition of humanity – that predisposition, as we have seen, in which Kant locates the radical evil in human nature.

“The predisposition to humanity can be brought under the general title of a self-love that is physical and yet involves comparison (for which reason is required); that is, only in comparison with others does one judge oneself happy or unhappy. Out of this self-love originates the inclination to *gain worth in the opinion of others*, originally, of course, merely *equal worth*: not allowing anyone superiority over oneself, bound up with the constant anxiety that others might be striving for ascendancy; but from this arises gradually an unjust desire to acquire superiority for oneself over others. – Upon this, namely, *jealousy* and *rivalry*, can be grafted the greatest vices of secret or open hostility to all whom we consider alien to us. These vices, however, do not really issue from nature as their root but are rather inclinations, in the face of the anxious endeavor of others to attain a hateful superiority over us, to procure it for ourselves over them for the sake of security, as preventive measure; for nature itself wanted to use the idea of such a competitiveness (which in itself does not exclude reciprocal love) as only an incentive to culture.” (R 6:27).

The original meaning of our natural desire for happiness is that we should compare our state with that of others and find it superior to theirs. As culture develops, our original defensive anxiety to protect ourselves against the ascendancy of others is transformed into a desire for superiority over them. This inclination does not issue directly from nature but arises from the development and use of reason, in setting ends and pursuing

happiness. When this competitive spirit is set alongside the basic requirements of the moral law – not to make an exception of ourselves to maxims we will to hold as universal laws, to treat all rational beings as ends in themselves rather than subordinating them to our ends, to follow the laws of a realm of ends, in which human ends are in systematic harmony – we see that it is in direct conflict with these moral demands (G 4:421-436). Once we see that our natural inclinations, when shaped by our social condition as rational beings, involve this competitive spirit, then we can see that the fundamental maxim of evil, which gives their satisfaction priority over obedience to the moral law, is really nothing except a desire for superiority over others, and a policy of esteeming ourselves on the basis of our state or condition, which can be compared with that of others with the aim of validating that superiority.

Much of Kant's working out of the theme of unsociable sociability in his historical, ethical and anthropological writings has to do with the various ways in which the self-esteem of individuals clashes, or in which people seek the three principal objects over which they compete – namely, power, wealth and honor. These, Kant says, are the means by which we hope to dominate others, making use (respectively) of their fear, their self-interest and their opinion (G 4:393, VA 7:271-273). But his reference to our “hostility toward all whom we consider alien to us” is significant, in that it implies a collective dimension to unsociable sociability – encompassing national, ethnic or religious forms of hostility between people. Kant points out that religions frequently invoke the power of deities on behalf of one nation or faith in its combat with others, and use alleged divine favor as a pretext for claiming dominance for their group over another (VpR 28:1124-1125). And of course it is war between nations that Kant regards as the form of evil that,

at this stage of human history, poses the greatest obstacle to the further progress of the human species (EF 8: 360-368; I 8: 24-27).

Kant is even more explicit about the social origin of evil at the beginning of Part Three of the *Religion*, where his aim is to show that the struggle against evil cannot succeed so long as each of us fights the moral battle apart from others, but has a chance of success only when people join together in an ethical community, taking the highest good as a shared or collective end and recognizing the moral law as a public (though non-coercive) law.

“If [the human being] searches for the causes and the circumstances that draw him into this danger [of subjection to the evil principle] and keep him there, he can easily convince himself that they do not come his way from his own crude nature, so far as he exists in isolation, but rather from the human beings to whom he stands in relation or association. It is not the instigation of nature that arouses what should properly be called the *passions*, which wreak such great devastation in his originally good predisposition. His needs are but limited, and his state of mind in providing for them moderate and tranquil. He is poor (or considers himself so) only to the extent that he is anxious that other human beings will consider him poor and will despise him for it. Envy, addiction to power, avarice, and the malignant inclinations associated with these, assail his nature, which on its own is undemanding, *as soon as he is among human beings*. Nor is it necessary to assume that these are sunk into evil and are examples to lead him astray; it suffices that they are there, that they surround him, and that they are human beings, and they will mutually corrupt each other’s moral disposition and make one another evil” (R 6:93-94).

This is as clear a statement as one could ask for that the radical evil in human nature arises and manifests itself only in the social condition. For Kant it is also only in the social condition that our reason is capable of developing, so it is also only in society that people could come to awareness of the moral law and could recognize evil for what it is. So it is human society which constitutes the condition both for evil and for the moral struggle against it. This same vision of the human predicament is present in all Kant’s writings on human history, for example, in the Fourth Proposition of *Idea for a Universal History with a Cosmopolitan Aim* (1784):

*The means nature employs in order to bring about the development of all their predispositions is their **antagonism** in society, insofar as the latter is in the end the cause of their lawful order.* Here I understand by ‘antagonism’ the *unsociable sociability* of human beings,<sup>10</sup> i.e. their propensity to enter into society, which, however, is combined with a thoroughgoing resistance that constantly threatens to break up this society. The predisposition for this obviously lies in human nature. The human being has an inclination to become *socialized*, since in such a condition he feels himself as more a human being, i. e. feels the development of his natural predispositions. But he also has a great propensity to *individualize* (isolate) himself, because he simultaneously encounters in himself the unsociable property of willing to direct everything so as to get his own way, and hence expects resistance everywhere because he knows of himself that he is inclined on his side toward resistance against others. Now it is this resistance that awakens all the powers of the human being, brings him to overcome his propensity to indolence, and, driven by ambition, tyranny and greed, to obtain for himself a rank among his fellows, whom he cannot *stand*, but also cannot *leave alone*. Thus happen the first true steps from crudity toward culture, which really consists in the social worth of the human being; thus all talents come bit by bit to be developed, taste is formed, and even, through progress in enlightenment, a beginning is made toward the foundation of a way of thinking which can with time transform the rude natural predisposition to make moral distinctions into determinate practical principles and hence transform a *pathologically* compelled agreement to form a society finally into a *moral whole*” (I 8:20-21).

Another Kantian name, therefore, for the radical evil in human nature is “unsociable sociability” – the need that human beings have not as merely animal beings but as rational beings for society with others, which, however, is also a need to gain superiority over them in honor, power and wealth. Unsociable sociability makes human society the scene of inequality and conflict – all the more so, at least up to this stage of history, insofar as human beings have become cultivated and civilized, and their rational powers have developed through the prodding offered by this same social competitiveness. The paradoxically two-faced propensity of unsociable sociability is what Kant means when, in the religion, he speaks of human competitiveness as “not excluding mutual love”, and the natural purposiveness of human competition is what he means when he says that nature employs human competition “as only an incentive to culture” (R 6:27).

These claims should be understood in the context of Kant’s theory of natural teleology in the study of living things, and its application to the history of the human

species. Biology and human history present us with phenomena that are governed by causal laws, but they also involve a kind of intelligibility which escapes explanation through these laws. Our best access to this intelligibility is through the regulative employment of the idea of an organized being, and of species of such beings, each with its own distinctive set of natural predispositions, for whose complete development nature has arranged. Kant's fullest explanation of this theory is found in the second half of the *Critique of the Power of Judgment* (KU 5:359-484). In a rational species this involves a *historical* process, in which each generation receives the skills and faculties developed by previous generations and then develops them further. The competitiveness of the social condition, which is made possible by the human propensity to evil, or unsociable sociability, is the natural mechanism through which this natural development takes place. Evil is therefore intelligible (to the extent that it can be made intelligible at all) as a mechanism employed by natural purposiveness in developing our species predispositions in history.

In *Idea for a Universal History*, however, Kant regards the era of history in which unsociable sociability can serve the historical development of human nature as having reached its limit. There he argues that it can continue to do so only if the human tendency to injustice is held in check by "a civil society universally administering right" – that is, a political state coercively enforcing laws of justice (I 8:22-23). This, in turn, will increasingly depend on the capacity of the human species to achieve a federation of political states maintaining peace with justice between them (I 8:24-26, cf. EF 8:360-368). Part Three of the *Religion* develops this thought further, by arguing that the moral progress of the human species depends on a different kind of human community, an

ethical community, grounded on non-coercive moral laws, in principle embracing the entire human species as “a people of God, and indeed in accordance with the laws of virtue” (R 6:99). The social model here is rather one of friendship, or of a family “under a common but invisible moral father...a free universal and enduring union of hearts” (R 6:102). It is this same vision, though stated in more secular terms, that Kant presents at the very conclusion of his *Anthropology from a Pragmatic Point of View*. There he is attempting to articulate the “character” of the human species as a whole, which he says can be done only historically, in terms of its moral vocation. We do this best in that judgment of our species which condemns it for its evil, which judgment also reveals in us the predisposition to good.

“So it presents the human species not as evil, but as a species of rational beings that strives among obstacles to rise out of evil in constant progress toward the good. In this its volition is generally good, but achievement is difficult because one cannot expect to reach the goal by a free agreement of individuals, but only by a progressive organization of citizens of the earth into and toward the species as a system that is cosmopolitically combined” (VA 7:333).

Kant’s vision of human history as proceeding by way of unsociable sociability toward a future moral unity is obviously inspired by Rousseau’s vision of the human species in his *Discourse on the Origin of Inequality*.<sup>11</sup> In Rousseau too, the development of our rational faculties occurs only in conjunction with the rise of social competitiveness, conflict and inequality. The development of reason transforms our innocent, animal self-love or *amour de soi* into a new impulse, which Rousseau calls *amour propre*. This distinction is reproduced in Kant’s moral psychology, with his contrast between “self-love” (*Eigenliebe*) and “self-conceit” (*Eigendiinkel*) (KpV 5:73) -- making “self-conceit”, along with “unsociable sociability,” yet another Kantian name for the radical propensity to evil.

Kant agrees with Rousseau's pessimistic assessment of the results of developing our rational faculties if one considers only what human beings have made of themselves so far in their history. The point, however, he thinks, is to view our species as faced with the challenge of making of itself something that it never yet has been, realizing the promise of our moral vocation by achieving the ends morality sets for us: "Rousseau was not so wrong when he preferred to it the condition of savages, as long, namely, as one leaves out this last stage to which our species has yet to ascend" (I 8:26).

In this manner one can also bring into agreement with themselves and with reason the assertions of the famous *J.-J. Rousseau*, which are often misinterpreted and to all appearance conflict with one another. In his writing *on the influence of the sciences* and *on the inequality of human beings*, he shows quite correctly the unavoidable conflict of culture with the nature of humankind as a *physical* species in which each individual was entirely to reach his vocation; but in his *Émile*, his *Social Contract* and other writings, he seeks again to solve the harder problem of how culture must proceed in order properly to develop the predispositions of humanity as a *moral* species to their vocation, so that the latter no longer conflict with humanity as a natural species. From this conflict (since culture, according to true principles of *education* of human being and citizen, has perhaps not yet rightly begun, much less having been completed) arises all true ills that oppress human life, and all vices that dishonor it; nevertheless, the incitements to the latter, which one blames for them, are in themselves good and purposive as natural predispositions, but these predispositions, since they were aimed at the merely natural condition, suffer injury from progressing culture and injure culture in turn, until perfect art again becomes nature, which is the ultimate goal of the moral vocation of the human species" (MA 8:116-118).

**Three objections.** The centrality of this Kantian conception of human history in Kant's anthropological and ethical writings is plain enough to anyone who has taken the trouble to become familiar with them. The evidences of the same vision in the *Religion's* account of radical evil are, as we have seen, equally clear and explicit, even if Kant's purpose in the *Religion* keeps him from giving them, at least until Part Three, the prominence they might seem to deserve. In several earlier writings, I have tried to emphasize the social and historical context of evil in Kant's account, but have

encountered several objections to the thesis that Kant regards the social condition as the context of radical evil.<sup>12</sup> This is perhaps a good place to reply briefly to three of them.

*Objection 1: Intelligible freedom.* One objection has been that to see the social condition as the context of radical evil is inconsistent with Kant's doctrine that we are free beings only in the intelligible world. This objection is based, in my view, on some very fundamental errors about Kant's treatment of the problem of freedom and the role of transcendental idealism in resolving it. The function of Kant's idea that we might be free as members of the intelligible world is only to show that there is no contradiction in regarding our actions both as free and as subject to the causal mechanism of nature in the sensible world (see KrV A557-558/B585-586).<sup>13</sup> Nothing Kant says could justify ascribing to him the absurd metaphysical fantasy that as free agents we are locked away in little monastic cells somewhere up there in the noumenal world. All Kant's writings on history and anthropology confirm the contrary intention to understand our moral condition in a natural, social and historical context. It is certainly common enough for Kant to be accused of taking a view of our free agency that detaches it entirely from every natural, social or historical context. Kant's view that evil develops in us only as social beings and as part of a natural teleology in human history certainly gives the lie to these sadly common misconceptions. Those who raise this objection show only that they prefer the common errors to the plain truth.

*Objection 2: Duties to oneself.* A second objection has been that to ascribe radical evil to the social condition accounts only those forms of evil that involve the violation of duties to others, and cannot encompass the violation of duties to ourselves. But this involves the failure to realize that in the social condition what happens to us is, most

fundamentally, that we come to value ourselves in the wrong way, preferring the worth of our condition – which can be compared favorably to that of others – over the worth of our person, which is measured not by comparison with others but only by the moral law (MS 6: 435-436, KpV 5:76-77, VE 27:349). The ground of all violation of duties to oneself is the failure to respect one’s own worth as a rational being, and this failure is most fundamentally what manifests itself in unsociable sociability and the kind of self-conceit that goes along with it. Besides, a closer look at Kant’s discussion of our self-regarding vices as moral beings -- lying, avarice and servility – shows all of them to be deeply social in their context and motivation. They are manifestations of social corruption every bit as much as our other-regarding vices (MS 6:429-437, VE 27:399-405, 604-607).

*Objection 3: “Blame society, not the individual.”* The third objection is that ascribing the radical evil in human nature to our social condition involves making society, or other people, responsible for our evil choices, which is inconsistent with Kant’s view that each of us alone is responsible for them.<sup>14</sup> But it is one thing to say that the social condition provides the necessary context for developing our radical propensity to evil, and quite another to say that society forces us to choose evil maxims, removing or diminishing our responsibility for these choices. Kant’s assertion of the first thing is plain in his texts, but he never says the second. If the objectors think that the one commits him to the other, then they are criticizing Kant, not interpreting him.

When Kant says that in the social condition human beings “mutually corrupt each other’s moral disposition and make one another evil” (R 6: 94), he obviously means that the presence of other human beings plays a necessary role our choice of evil maxims, but not that their influence provides us with any excuse for our wrong choices. Kant

understands “a corrupted disposition” as something for which *the corrupted person* is morally responsible -- otherwise it would not count as *moral corruption*. When we speak, on a less fundamental level, of bad education or bad company as “corrupting” a person’s character, we mean that they play a role in making the person corrupt or wicked, not that they release the person from responsibility for their bad character. No doubt a bad social environment can sometimes serve to excuse or even justify someone’s conduct, which would otherwise deserve blame. But this is plainly not what Kant has in mind, and it would be deplorably simple-minded to think that all harmful social influences must exculpate bad conduct. We saw above how Kant is falsely accused of removing free ethical choice from its natural and social context. But now it is not Kant but the objector who thinks that admitting any context at all for free choices destroys their freedom and undermines the chooser’s responsibility.

Or is the objection that the unsociable sociability of others puts me under overwhelming pressure to treat them badly, perhaps as my only defense myself against their bad conduct? But this is plainly not how Kant views the matter. My proper defense against others is to demand justice and respect from them, which is not to do evil. It is not to treat them with cruelty and deceit, injustice and hatred, which is to do evil. Rivalry with others does not justify my bad conduct but rather constitutes only a temptation to evil. To offer the fact that I was tempted as an excuse for bad conduct is a pitiful ploy, inviting derision as well as added blame. Obviously the moral law commands me to resist such temptations, and Kant frequently insists that my predisposition to personality gives me the capacity to do so (R 6:49).<sup>15</sup>

Kant's account does explain evil *teleologically* by showing how it serves the natural purpose of developing our species predispositions (I 8:22-24). Does this explanation contradict the claim that we are responsible for evil? Kant does not think we can provide anything like a causal explanation for our choice of evil, describing this choice as "inscrutable to us" (R 6:21). It is a basic misunderstanding of Kant's conception of inner natural teleology to think that it operates through some external agent ("God" or "nature") causally determining what happens. Kant's account is *not* that "nature" causes us to make evil choices using "society" as its causal mechanism.

**Conclusion.** Kant's account makes evil intelligible in two ways: first, by identifying the fundamental maxim of evil, and second, by locating our propensity to evil within the context: of our social condition, and the natural teleology in its history. Reason, morality and the propensity to evil are all fruits of our social condition, and in that condition our vocation is to employ our reason, and the principle of morality that it reveals to us, in the struggle against all the evil propensities of self-conceit, unsociable sociability, tyranny, greed and ambition – which, however, served as the historical conditions for the development of reason and even of morality itself. Kant's attitude toward the likelihood of our success in this struggle is hardly one of confident optimism, but it is one of sober, principled hopefulness.

Nowadays we may not find Kant's theory of natural teleology a persuasive setting in which to place either a conception of human history or an attempt to make evil intelligible as a part of it. It would be more fashionable for someone to speculate that our propensity to seek superiority over others, along with the competitiveness this entails and the development of human capacities that results from it, belong to traits that were

selected for early in the evolution of the human species during the period when it was socially organized into troops of hunter-gatherers on the plains of Africa. Evolutionary biology has explained many things and may explain still more, but when it comes to the contingencies of human history, I think unconfirmable Darwin-inspired just-so stories are seldom any improvement over pre-Darwinian speculations they were designed to replace. Kant's theory of history, based on a cautious natural teleology grounded on a theory of reflective judgment, may still be as good as anything we have so far come up with.

Kant's account of evil locates evil itself, as well as our struggle against it, in our forms of social organization and their history. It treats the development of human reason as part of our cultural history, and evil as a vehicle in this development, including the genesis of our ability to recognize evil for what it is and to struggle against it. Evil is grounded in antagonistic or competitive social relations, and while the struggle against it is fundamentally a striving to strengthen human solidarity and bring human purposes into that systematic agreement to which Kant gave the name "the realm of ends". Kant's attempt to make evil intelligible shows us that society and social change are the framework in which we ought to think about both the sources and the remedies of the evil people do.

## Notes

---

<sup>1</sup> Kant's writings will be cited according to the following system of abbreviations:

- Ak *Immanuel Kants Schriften*. Ausgabe der königlich preussischen Akademie der Wissenschaften (Berlin: W. de Gruyter, 1902-). Unless otherwise footnoted, writings of Immanuel Kant will be cited by volume:page number in this edition.
- Ca *Cambridge Edition of the Writings of Immanuel Kant* (New York: Cambridge University Press, 1992-) This edition provides marginal Ak volume:page citations. Specific works will be cited using the following system of abbreviations (works not abbreviated below will be cited simply as Ak volume:page).
- EF *Zum ewigen Frieden: Ein philosophischer Entwurf* (1795) , Ak 8  
*Toward perpetual peace: A philosophical project*, Ca Practical Philosophy
- G *Grundlegung zur Metaphysik der Sitten* (1785), Ak 4  
*Groundwork of the metaphysics of morals*, Ca Practical Philosophy
- I *Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht* (1784), Ak 8  
*Idea toward a universal history with a cosmopolitan aim*, Ca Anthropology History and Education
- KrV *Kritik der reinen Vernunft* (1781, 1787). Cited by A/B pagination.  
*Critique of pure reason*, Ca Critique of Pure Reason
- KpV *Kritik der praktischen Vernunft* (1788), Ak 5  
*Critique of practical reason*, Ca Practical Philosophy
- KU *Kritik der Urteilskraft* (1790), Ak 5  
*Critique of the power of judgment*, Ca Critique of the Power of Judgment
- MA *Mutmaßlicher Anfang der Menschengeschichte* (1786), Ak 8  
*Conjectural beginning of human history*, Ca Anthropology History and Education
- MS *Metaphysik der Sitten* (1797-1798), Ak 6  
*Metaphysics of morals*, Ca Practical Philosophy
- R *Religion innerhalb der Grenzen der bloßen Vernunft* (1793-1794), Ak 6  
*Religion within the boundaries of mere reason*, Ca Religion and Rational Theology
- SF *Streit der Fakultäten* (1798), Ak 7  
*Conflict of the faculties*, Ca Religion and Rational Theology
- VA *Anthropologie in pragmatischer Hinsicht* (1798), Ak 7  
*Anthropology from a pragmatic point of view*, Ca Anthropology, History and Education  
*Vorlesungen über Anthropologie*, VA 25  
*Lectures on Anthropology*, Ca Lectures on Anthropology
- VE *Vorlesungen über Ethik*, Ak 27, 29  
*Lectures on Ethics*, Ca Lectures on Ethics
- Vpr *Vorlesungen über die philosophische Religionslehre*, Ak 28  
*Lectures on the philosophical doctrine of religion*, Ca Religion and Rational Theology

---

<sup>2</sup> This metaphor is drawn from Gordon E. Michalson, *Fallen Freedom: Kant on radical evil and moral regeneration*: Cambridge, Eng.: Cambridge University Press, 1990.

<sup>3</sup> Even if we grant the rationalist assumption that there are decisive reasons against doing evil actions, it may still be true that evildoing is often “rational” in some perfectly obvious acceptable sense – an evil action may, for instance, be the best means available to the thing the agent wants most (such as that agent’s own happiness). Further, even if evil is always contrary to reason, because there are always decisive reasons for not doing it, we may still grant that evil, or at least a lot of it, should not be considered “irrational”. For we are not in the habit of applying that term in cases where the reasons someone is acting against are clearly known to the agent but the agent simply refuses to see them as reasons. We who do recognize the rejected reasons, however, must still judge the agent open to criticism on rational grounds.

<sup>4</sup> Deliberate evil, unlike self-deception and *akrasia*, probably does not count as a case of ‘irrationality’ in the usual sense of the term, where we call thinking or behavior ‘irrational’ only if it runs contrary to reasons or standards of rationality that the agent explicitly accepts, or would accept (if the issue were put to them). We do not usually call ‘irrational’ the deliberate refusal to act according to the best reasons one has, or to recognize them as valid reasons at all. (But I think if we regarded the reasons as obvious enough, we might treat this too as a case of irrationality; so it says something about *us* – probably something pretty unflattering -- if we do not see the overridingness of moral reasons as obvious). This point about the ordinary usage of ‘irrational’ has been taken by some (e.g. by Bernard Williams) to call into question whether there are any genuine reasons applicable to an agent that the agent does not acknowledge (stigmatizing these as ‘external’ reasons). It is therefore sometimes taken to be an argument for the substantive thesis that we have no reason to act as morality requires unless we acknowledge such reasons. But if it is true that morality provides us with reasons to meet its requirements, then conduct that refuses to recognize such reasons, though perhaps not ‘irrational’, does nevertheless exhibit a clear failure of rationality, and it is open to rational criticism. For a good discussion of this point, see T. M. Scanlon, *What we owe to each other* (Cambridge, MA: Harvard University Press, 1998), pp. 25-30.

<sup>5</sup> This is not to deny that we can inquire into the kinds of situations in which people are likely to make evil choices, perhaps with a view to avoiding those situations and thus avoiding evil and its effects. This is a point sometimes emphasized by ethical ‘situationists,’ such as John Doris and Gilbert Harman. They are certainly correct that it is important to know what situations these are, and to do what we can to prevent them. The point, however, is that these situations do not *cause* evil choices (what a situation could cause would not be *evil*, but only some event, to which we might take a negative attitude) but only provide the occasion for human beings to make them, or for evil propensities in people to show themselves.

<sup>6</sup> See John R. Silber, “The Ethical Significance of Kant’s *Religion*,” in Theodore M. Greene and Hoyt Hudson, *Kant, Religion within the Limits of Reason Alone* (New York: Harper, 1960), pp. cxxv-cxxvii, and Richard Bernstein, *Radical Evil: A Philosophical Interrogation* (Cambridge, MA: Polity Press, 2002), pp. 36-42.

<sup>7</sup> For this reason, it is not clear that what is being described here as the ‘diabolical’ will could truly be an *evil* will at all. Elsewhere, however, Kant speaks of the vices of culture, though of in their “extreme degree, that surpasses humanity,” as ‘diabolical vices’ (R 6:27, cf. MS 6:461), which clearly are evil in the proper sense of the term. Here he seems to have in mind the vices of hatred – envy, ingratitude and malice, as well as the vice of rejoicing in others’ misfortunes. But in saying that in their extreme degree they “surpass humanity”, Kant means to discourage us from thinking of such extremes of evil as actually found in human beings, just as he does not encourage us to think of the contrary (‘angelic’) virtues as found in actual human beings (MS 6:458-461). The point in both cases, I think, is that we would do well not to project onto others either our resentment at vice or our admiration of virtue, but to concentrate instead on what we have in common with other human beings (in the way of both virtue and vice) and recognize both the best and the worst of our fellow human beings as not all that different from ourselves.

<sup>8</sup> Kant’s view at this point therefore involves (contrary to the mistaken claims of Bernstein, see note 5 above) no restriction on the scope of human freedom – no limitation on what human beings may choose. It

---

is rather a view about the structure of human incentives – a view about what they must be if any choice human beings make is rightly to be called ‘good’ or ‘evil’ at all.

<sup>9</sup> For a similar recent discussion, see Matthew Caswell, “Kant on the Diabolical Will: A Neglected Alternative?” *Kantian Review*, Volume 12-2 (2007), pp. 147-157. Caswell does a good job of making the point that viewing someone as so fundamentally evil in their motivations that they are *incapable* of good is not only morally incoherent but it is also to imagine their evil as “radically alien” to ours in a way which tends to blind us to what is truly evil, especially in ourselves (p. 156). Cf. “[Such views] are best explained as the projection of irrational hatred and resentment, which are not uncommon.” Allen Wood, *Kant’s Ethical Thought* (New York: Cambridge, 1999), p. 401.

<sup>10</sup> This phrase is taken from Montaigne: “Il n’est rien si dissociable et sociable que l’homme: l’un par son vice, l’autre par sa nature.” Michel Eyquem de Montaigne, “De la solitude,” *Essais*, ed. André Tournon (Paris: Imprimerie nationale Éditions, 1998), I:388. “There is nothing more unsociable than Man, and nothing more sociable: unsociable by his vice, sociable by his nature,” “Of Solitude,” *The Complete Essays*, tr. M. A. Screech (London: Penguin Books, 1991), p. 267.

<sup>11</sup> Jean-Jacques Rousseau, *Discourse on the Origin of Inequality*, tr. Donald Cress. Indianapolis: Hackett, 1992.

<sup>12</sup> For instance, see *Kant’s Ethical Thought* (New York: Cambridge University Press, 1999), pp. 283-320, and “Religion, Ethical Community and the Struggle against Evil,” *Faith and Philosophy* 17/4 (2000), pp. 498-511. My eyes were first opened to this theme in Kant’s anthropological, historical moral and religious thought by Sharon Anderson-Gold, “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, IN: Indiana University Press, 1991), pp. 113-131, as well as by discussions of this theme with her at the conference on which this volume was based.

<sup>13</sup> See also my paper “Kant’s Compatibilism,” in Wood (ed.) *Self and Nature in Kant’s Philosophy* (Ithaca: Cornell University Press, 1984), especially the final paragraph on p. 99: “In assessing Kant’s compatibilism, it may help to remind ourselves that his theory of timeless agency is put forward only as a means of exploiting the burden of proof in the free will problem, which falls on those who would show that freedom is incompatible with determinism. Kant is not positively committed to his theory of the case as an account of the way our free agency actually works. Indeed, Kant maintains that no such positive account can ever be obtained. Kant does not pretend to know how our free agency is possible, but claims only to show that the impossibility of freedom is forever indemonstrable. If what bothers us about Kant’s theory is that it seems too far-fetched and metaphysical, then it may help at least a little to realize that once the theory has served as a device for showing that freedom and determinism cannot be proven incompatible, he is just as content to dissociate himself from it and adopt a largely agnostic position on the question how our freedom is possible.” Many discussions of this article have proceeded as if I had never written these words, and have supposed I meant to defend Kant’s notion of noumenal freedom as a dogmatic metaphysical doctrine. If I had anticipated such profound misunderstandings, I would certainly have given this point more emphasis. But it is also true that the capacity of philosophers maliciously to misunderstand what other philosophers have written is virtually infinite, so perhaps nothing I could have done would have made any difference.

<sup>14</sup> Jeanine Grenberg, *Kant and the Ethics of Humility: A Story of Dependence, Corruption and Virtue* (Cambridge: Cambridge University Press, 2005), pp. 31-42.

<sup>15</sup> “Suppose someone asserts of his lustful inclination, when the desired object and the opportunity are present, it is quite irresistible to him; ask him whether, if a gallows were erected in front of the house where he finds this opportunity and he would be hanged on it immediately after gratifying his lust, he would not then control his inclination. One need not conjecture very long what he would reply. But ask him whether, if his prince demanded, on pain of the same immediate execution, that he give false testimony against an

---

honorable man whom the prince would like to destroy under a plausible pretext, he would consider it possible to overcome his love of life, however great it may be,. He would perhaps not venture to assert whether he would do it or not, but he must admit without hesitation that it would be possible for him. He judges, therefore, that he can do something because he knows he ought to do it, and cognizes freedom within him, which, without the moral law, would have remained unknown to him” (KpV 5:30).