Kant on the Diabolical Will: a Neglected Alternative?

MATTHEW CASWELL
St John's College, Annapolis

To his harshest critics, Kant's philosophy can seem an unending series of neglected alternatives. Time and again, Kant argues for his position by elimination, ruling out each possible alternative, until his own is the only one left standing. Of course, this strategy amounts to a demonstration of the Kantian position if and only if the field of possible alternatives really is – as Kant always assumes – exhaustive. But readers often suspect that Kant has stacked the deck, that his dogmatic adherence to a particular set of presuppositions, or perhaps simply his lack of philosophical imagination, has unjustifiably restricted the possibilities under consideration, thereby rendering the alleged argument by elimination nothing but an elaborate exercise in question begging.

Kant's various declarations that certain phenomena are strictly impossible seem almost designed to provoke the critical reader. *Never* another Newton of a blade of grass? Well, why not? The dialectical movement begun in moments like these drives us to explore possible Kantian responses to the inevitable 'Why not?' challenge, and this turns out to be a potentially powerful way to illuminate Kant's theory, and indeed to evaluate it. Kant offered a particularly provocative prohibition in his theory of evil elaborated in Book One of the *Religion (Rel.)*. Kant there asserts that evil for evil's sake is simply not possible for human beings: there is no 'diabolical will' for beings like us. According to Kant, we are certainly capable of freely, knowingly, disobeying the moral law, but we never repudiate morality 'in the manner of a rebel' (*Rel.* 6: 36), rejecting the law out of sheer hatred for virtue (or sheer love of evil). Rather, Kant argues, human evil is always done from self-love, and can only arise from a subordination of moral principle to our own happiness.

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

Kant's position here is of a piece with the moral–psychological theory developed in the *Religion*. In this text, Kant adds much to his standard account of the will as practical reason, as a faculty that enacts maxims, or subjective practical principles. Namely, Kant here argues that an agent's set of maxims has a single, overarching or ultimate maxim, governing the 'whole use of our freedom' (*Rel.*, 6: 25), guiding the adoption of all other subordinate maxims. This ultimate subjective principle constitutes the agent's *Gesinnung* or basic cast of mind. The *Gesinnung* is an 'ultimate subjective principle' in the sense that it justifies the adoption of all other maxims, and cannot be justified in light of some even higher-order maxim (*Rel.* 6:21n.). Therefore, an agent's adoption of any evil maxim entails that his ultimate justificatory ground for the adoption of maxims – his *Gesinnung* – licenses the violation of moral dictates, and is itself therefore evil. That is, evil in human agents always goes all the way down, to the root ground of our character.

In the willing of any maxim, an incentive to the will is freely incorporated into a rational principle, rendering it for the first time a motivational ground, or reason to act. This thesis is crucial for securing the freedom and accountability of human action. As Kant writes, freedom of the will is of a wholly unique nature in that an incentive can determine the will to an action only so far as the individual has incorporated it into his maxim (*Rel.* 6: 24).

Now, in the *Gesinnung* for each human agent, Kant argues that exactly two generic incentives are incorporated: that of self-love, which motivates the pursuit of happiness, and that of respect, which motivates morally worthy action from duty. Since neither of these incentives is in itself immoral, an evil *Gesinnung* could result only from the subordination of the moral incentive to that of self-love. Therefore, evil action is always grounded in a self-love made the overriding condition of action, not in a love of evil for its own sake.

The first objection Kant's prohibition on diabolism is likely to provoke in readers is a phenomenological one. That is, we might counter that Kant's own position is simply incapable of explaining certain all-too-well-known behaviours. Some knowingly evil agents pursue their malevolent projects at great personal risk and sacrifice – they are apparently not driven by the allure of happiness, but are even selflessly evil.[1] The suicidal terrorist offers a familiar case of disciplined, self-denying murderousness. Objectors might also

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

point out a clear 'wickedness gap' between the commonplace self-preference of the evil agent in Kant's account and the monstrous horrors of, for example, the last century.

Kant in fact has a sophisticated response available to such objections. As has been explored by Henry Allison[2] and Sharon Anderson-Gold,[3] Kant's theory of 'unsocial sociability', according to which human happiness is always defined within a competitive social context, helps explain how ends of domination and subjugation can come to count as part of an agent's happiness, such that an evil agent might even sacrifice health and well-being for their sake. The prospect of acquiring the title 'martyr' can, thanks to the complicated social network through which we define our happiness, come to seem a more valuable part of our happiness than life itself. Thus, apparently selfless destruction can be plausibly seen as motivated by concern for a highly socialized conception of one's personal status, and not out of a diabolical love of evil for evil's sake.[4]

But this Kantian rebuttal is only partially successful. Let us grant that a Kantian armed with Kant's unflattering theory of human nature *can* give a plausible account of the suicidal terrorist, for example, or the Nazi decision to draw back the war effort – imperilling victory – in order to murder more Jews at a faster rate, *without* attributing to them a diabolical will. But even if the Kantian's explanation is plausible, how can we be sure these criminals were not diabolical after all? The Kantian response thus far amounts merely to the thesis that the conception of a diabolical will is not *necessary* to explain human behaviour. It does not appear to follow from this thesis alone that it is *impossible* to attribute a diabolical will to beings like us. How can this stronger thesis be defended?

Kant's prohibition on diabolical willing is in fact a claim about the necessary conditions of accountable immorality. The most fundamental claim of Kant's moral theory is perhaps what has become known as the Reciprocity Thesis: that the unconditional obligation imposed by the moral law and the strict freedom of a will so obligated reciprocally entail one another.[5] This link between freedom and responsibility applies to the good will as much as the evil: to count as evil, an agent must be free, and obligation to the moral law provides the only possible grounds for inferring that a will is free. Given this background thesis, Kant argues that to 'think of oneself as a freely acting being, yet as exempted [*entbunden*] from the one law commensurate to such a being (the moral law), would amount

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

to the thought of a cause operating without any law at all ... and this is a contradiction' (*Rel.* 6: 35).

That is, we cannot conceive of a diabolical will 'exempted' or 'released' from the moral law without sacrificing such a will's freedom, and such an unfree will, which could not be held accountable for its conduct, could not qualify as evil. Kant's strong claim that the notion of a diabolical will is a 'contradiction' must be understood to turn on the reciprocal entailment of moral obligation and freedom. The contradiction lies in the fact that such a will – qua evil – is thought to be free, and at the same time is thought – qua exempt from moral law – as not subject to the necessary conditions of freedom. Kant is not claiming that the very idea of a non-moral, non-natural causal power contradicts itself; rather the attribution of the thoroughly moral predicate 'evil' to such a power is contradictory. We may indeed confront agents in the world, such as the classic psychopathic criminal, who we have reason to believe violate morality for its own sake. But these agents are exempt from the moral law, incapable of legislating it for themselves, and therefore do not merit moral blame: they are sick, not evil. Thus, a diabolical will could be seen as dangerous, as something to be avoided or destroyed, but not something admitting of moral condemnation.

This line of argument may be convincing enough on its own terms, but appears to leave open a neglected alternative.[6] Kant's model of diabolical will is of an agency operating outside of our own moral universe. It is motivated neither by inclination (which would make it self-interested) nor by pure practical reason (which would make it moral). Because it is never subject to the call of duty, such a will is not subject to temptation in the normal sense. Such a will would not experience remorse, moral struggle or inner conflict of any sort, since by hypothesis the love of evil holds sway unchecked by the moral incentive. This model of diabolism can be specified: call it the demonic will. Such an agent is truly an alien, whose principles and motives must remain for us thoroughly unintelligible and unaccountable. But this is the only model of diabolical willing available?

Richard Bernstein delineates an alternative model of diabolism, conceding that, while the human species as such cannot be judged as diabolical, Kant has not ruled out the possibility that individual agents, subject to law, might '*become* diabolical', freely subordinating morality to a love of evil for its own sake (38–9).[7] Indeed, it

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

seems quite possible to imagine an agent who chooses evil for its own sake, and yet does not conceive himself as exempt from the moral law. Unlike the demonic agent, this alternative agent – whom we may call the contrarian – would be obligated by morality, would struggle with its commands, and so could be taken as both free and accountable, but nevertheless would violate moral principle out of a love for evil as such.

These two models of the diabolical will – demonic and contrarian – can be contrasted relative to one another, and to the Kantian model of the human will, according to the incentives available for incorporation into their principles. First consider the incentives Kant allows as available to the human will. According to Kant, the motivational dualist, these include two and only two: morality and self-love, given through reason and inclination respectively. The demonic will, whatever unintelligible, lawless incentives it might be subject to, is lacking the moral incentive – it is 'exempt'. The contrarian model of the diabolical will can be thought of as 'dualism plus'. In order to maintain that we are free human agents, the advocate of the contrarian will grants that our *Gesinnung* incorporates both of Kant's dual incentives. But, as Bernstein asks, 'Why should we limit incentives to these two kinds?' (42). Couldn't a will incorporate both its concern for its own happiness and its respect for the moral law as subordinate to some third incentive, not reducible to either of the others? This third motivational ingredient would be, following Bernstein, an 'incentive to *defy* the moral law' (41). If so, the contrarian diabolist, unlike his demonic cousin, would indeed subordinate morality. That is, the contrarian's *Gesinnung* would involve a tripartite ordering: love of evil is its highest priority, and the condition under which all other pursuits are promoted; either self-love or else morality assume second place, leaving the other to occupy third place. (Of course, advocates of a contrarian diabolical will might welcome a host of 'extra' incentives, perhaps even an infinite number of them.[8]) It seems clear from the text that Kant neglected this alternative. After cataloguing the incentives available to the human will, Kant points out that the presence of the moral incentive is a necessary condition of freedom, thereby ruling out the possibility of what we have titled the demonic model of diabolism. However, the contrarian will we have considered chooses evil for its own sake, and yet is not free from moral obligation. Because this agent is not 'exempt' from morality, its possibility appears to

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

be immune from the objection Kant brings against diabolism. In his account, Kant ignored such a conceptual possibility. But did he have to? Does the contrarian model amount to a neglected alternative for diabolical willing? Or are there Kantian grounds for ruling out the contrarian model of diabolical willing as well, thereby securing the prohibition on diabolism as such?

In fact, careful consideration of the logic of rational agency reveals that the contrarian will, so described, is not possible. The problem is that it is impossible to subordinate morality to its opposite, for it is impossible to subordinate any interest to an interest in its opposite. Consider a non-moral example. I may have competing interests in my career and in golf. Now, it is possible for me to subordinate my career interest to my golf interest, neglecting professional duties in order to spend more time golfing; or, alternatively, I could make my job my higher priority, and play golf only when my work schedule allowed. But it is *not* possible to subordinate my love of golf to a hatred of golf. If I did hate golf, and thus had an interest in avoiding golf whenever possible, I could not pursue this interest in a way that accommodated my love of golf, or my interest in playing golf whenever possible. These two interests (golf-love and golf-hatred) cannot share a home in the same will; either one is present or the other, but they cannot be subordinated relative to one another. We can elucidate this point schematically. Any maxim in which one interest (x) is subordinated to another (y) may be explicated as follows:

1. I will pursue x except on the condition that it interferes with my pursuit of y.

Substituting 'golf' and 'career' for x or y would generate a coherent policy. However, the following is not a coherent type of policy:

2. I will pursue x except on the condition that it interferes with my pursuit of not-x.

The adoption of such a policy would mean that an agent, for example, would take every opportunity to play golf except in those cases where he could take the opportunity not to play golf. Clearly, this is no subordination of golf-love to golf-hatred, but golf-hatred pure and simple. The allegedly subordinated interest in such a maxim cannot possibly be action-guiding, since the contradictory interest

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

will override it in every possible case. Thus, the agent's effective practical principle is nothing more than: I will always avoid golf-playing. The lesson here is that the finite rational will may arrange competing – or, as it were, *contrary* – interests such that some are subordinated to others in a hierarchy of priorities. In that case, the subordinated contrary interest will still guide action in cases where the higher priority allows. But the will cannot combine *contradictory* interests into a single system of maxims. Whichever of the two contradictories is 'subordinated' in fact vanishes, and cannot play any role in determining action.

This general point about maxims applies to the moral case of the ultimate maxim. Because the contrarian allegedly subordinates his respect for the moral law (and thus his interest in adhering to duty and pursuing virtue) to his hatred of morality, or his interest in violating moral dictates, his ultimate maxim must take something like the following form:

3. I will do whatever morality demands, except on the condition that doing so interferes with me defying the demands of morality.

That is, love of morality is subordinated to love of evil. But according to the rule elaborated above, this is not a possible maxim. For such a policy clearly reduces to the following:

4. I will defy the demands of morality.

Why couldn't the diabolical will have an ultimate maxim like this one (4)? In so far as this *Gesinnung* is devoid of the moral incentive, it characterizes a demonic will, 'exempt' from the moral law. Above, we saw why a demonic will cannot be attributed to human evil agency, since it could not be thought of as free. Thus, once the alleged subordination of virtue to evil has been uncovered as reducing to the willing of evil *simpliciter*, the neglected alternative posed by the contrarian diabolical will is revealed to be no alternative at all. Contra Bernstein, it is not possible for human agents to incorporate into their *Gesinnung* an 'incentive to *defy* the moral law' (41).[9]

Diabolical willing can be excluded, on this view, through a consideration of the constraints reason imposes on free willing as such.

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

It is crucial to note that these include what Kant calls 'analytic' constraints. The principle elucidated above, that the will cannot subordinate one interest to its contradictory, is just such a constraint, and should be seen on par with the instrumental principle discussed in the *Groundwork*, requiring that we take the necessary means to those ends we will. Neither of these are imperatives, but rather stipulate constraints 'already thought' in the act of willing. That is, it is not the case that one *ought* to conform to the analytic principles of practical reason, for we are never in a position to make that sort of choice. Rather, 'in so far as he is a rational agent' (*Gr*, 4: 417), an agent necessarily conforms. In so far as he is not a rational agent, his conduct is ethically meaningless.[10]

Kant's position, which places irrationality below good and evil (as it were), will certainly provoke objections from readers accustomed to thinking of irrationality as the hallmark of human immorality. It's worth noting that the term 'irrationality [*Irrationalität*]' never appears in Kant's writing.[11] For Kant, the relevant term of contrast with 'rational' is 'non-rational [*unvernünftig*]', as in the designation of animals as non-rational beings. Irrationality as an ethical problem, perhaps an inheritance from nineteenth-century thought, is quite foreign from Kantian moral theory. The human moral task is not, according to Kant, one of rendering our conduct rational, but rather one of bringing our conduct into conformity with reason's highest principle. In fact, a close consideration of Kant's ethics may require us to ask whether the familiar notion of an irrational will is itself coherent.

It might be objected that Kant does attribute weakness or frailty to the human will. Isn't this an admission that the human will can conduct itself irrationally, and indeed that an important ethical phenomenon should be understood in these terms? Certainly, many thinkers take weakness of the will or akrasia as a paradigm example of ethically meaningful irrationality. But implicit in Kant's own treatment of evil lies a deep challenge to this way of framing the issue of akrasia. For Kant, moral weakness must be a self-imposed debility: that is, we choose our own frailty from an evil *Gesinnung*. A good dose of evil's trademark self-deception allows us to view the flagging of our moral efforts as a victory of desires over our good will, when in fact such a victory could never have been possible without 'a secret league' between our fallen will and the things that tempt it. The non-moral cases of alleged akrasia are a bit more com-

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

plicated. In Kant's view, prudential ends are not given us by nature, but are ends we give ourselves by freely incorporating them into our maxim (*Rel.* 6: 24). Accordingly, a putative case of prudential weakness – say, a failure of resolve in sticking to one's diet – would need to be reinterpreted as success in carrying out one's perfectly resolute (though perhaps only implicit) policy of enjoying fine pastries, at least in so far as responsibility for the akratic action was assigned to the agent. In both the moral and prudential cases, Kant can save the phenomena without recourse to a notion of an irrational will, and thus without placing those phenomena outside of the realm of the culpable.

Kant's concern to exclude the possibility of diabolical willing is part of his effort to find conceptual space for immorality within the community of moral persons. The motivation for such a concern is clear: outside of the moral community, praise and blame are inapplicable. Moreover, in Kant's view, implicit in the judgement that a will is evil is the notion that it ought not to be. Therefore, every evil will, because it is obligated to be good, must be capable of conversion and improvement. The diabolical agent, who might listen to our discussion of respect and duty with casual interest, could not act from such grounds, and therefore could not convert. The universality of moral principle entails that we all, good and evil, belong to a single moral community, subject to the same norms of moral judgement, and capable of living up to the same demands of duty.

To my mind, this is a decisive advantage of Kant's theory of immorality. For a thinker like Aristotle, the vicious agent is a lost cause, lacking the necessary practical wisdom to convert to (or even improve towards) virtue. An Aristotelian virtue ethics presents a moral universe segmented by a barrier between the knowing virtuous ones and the ignorant others. The same set of moral demands and judgements cannot extend across that barrier. In a related sense, a utilitarian also sees those who reject morality as acting according to a mistaken preference regarding the best kind of life, and therefore as incapable of choosing virtue over vice for themselves.

Towards the close of Kant's discussion of evil, Kant makes a striking plea for the egalitarianism of his moral universe. Attesting to the accessibility of the moral feeling of respect to 'even the man of greatest limitations', Kant recommends the use of 'examples of good men' in moral education. 'But,' he warns,

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

> to teach a pupil to admire [*bewundern*] virtuous actions, however great the sacrifice these may have entailed, is not in keeping with preserving his feeling for moral goodness. For be a man never so virtuous, all the goodness he can ever perform is still his simple duty; and to do his duty is nothing more than what it is in the common moral order and hence in no way deserving of wonder. Such wonder is rather a lowering of our feeling of duty, as if to act in obedience to it were something extraordinary and meritorious. (*Rel.* 6: 48–9)

The idea at work here is that of the moral equality of all members of the moral universe. We are all obligated by the same principles, and are all just as capable of fulfilling them. Few things could be worse for ethical development than if we came to see duty as something binding on and practicable by only an elite.

Although Kant does not explore it, an interesting parallel in the realm of immoral action suggests itself here. Kant would surely warn us not to romanticize the evil of those we condemn (ourselves included). If we think of evil agents as diabolical villains, with our condemnation we at once expel them from our moral community. They become radically different from us, acting from wholly alien incentives we must find unintelligible. As we have seen, such a conception of evil undermines the sense in which it can serve in moral condemnation; for the diabolical will is not free to be otherwise. Moreover, regarding evil with something akin to 'wonder' – perhaps horror – makes it all too easy to disregard our own propensities towards immorality. The barrier we construct in our conception of the moral universe between ourselves and other, allegedly diabolical, villains corrodes our own efforts at struggling against temptation, at becoming virtuous. There is a Kantian point to the cliché that we ought not demonize our enemies: to do so voids the ethical meaning of our moral judgement; and, worse, by imagining evil as radically alien, we put ourselves in a poor position from which to guard against it. Thus, the notion of a human diabolical will is not only conceptually incoherent on Kant's account, but ethically dangerous. Evil is not alien, argues Kant, but – no matter how grotesque – ordinary: criminals of every sort (at least the sane among them) belong to the same 'common moral order'. When judging the perpetrators of even the most horrible acts, we should struggle to keep the bounds of our moral community open.

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press

Notes

1 John Silber is the best-known spokesman for this objection, at least among Kantian commentators. See 'The Ethical Significance of Kant's Religion', introduction to *Religion within the Limits of Reason Alone*, trans. T. M. Greene and H. H. Hudson (New York: Harper Torchbooks, 1960), pp. cxxix–cxxx.

2 'Reflections on the Banality of (Radical) Evil', in *Idealism and Freedom* (Cambridge: Cambridge University Press, 1996), especially pp. 176–7.

3 'Kant's Rejection of Devilishness: the Limits of Human Volition', *Idealistic Studies*, 14 (1984), 35–48.

4 It is helpful here to recall Kant's discussion in the *Groundwork* of the naturally sympathetic 'friend of man' (*Gr*, 4: 398), for there Kant makes it clear that 'selfless' action done 'from self-love' is not in the least paradoxical on his account.

5 The phrase is Allison's, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), pp. 39–40.

6 Kant's own argument against the possibility of a diabolical will has been clearly articulated by Allen Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999), pp. 400–1, and Henry Allison (*Idealism and Freedom*, p. 176) in response to Silber's embrace of the possibility of diabolism. However, since neither Kant nor his contemporary defenders address the contrarian alternative posed by Bernstein and discussed below, I believe the Kantian response must be supplemented with further argumentation.

7 *Radical Evil: a Philosophical Interrogation* (Cambridge, MA: Polity Press, 2002), pp. 36–42.

8 Bernstein critiques Kant for not taking into account an unspecified 'full range of incentives ... involved in the adoption of evil maxims' (42).

9 In addition to advocating the possibility of a diabolical or devilish will, Bernstein makes the surprising claim that 'there is no free choice unless there is the free choice to be ... devilish' (42). I fail to see what could possibly count as evidence for this strong claim.

10 See Bernstein's claim that 'there are no *intrinsic* constraints on what the *Willkür* can choose to do' (42). But in so far as the *Willkür* is a faculty of practical reason, it is indeed subject to the intrinsic – as Kant would say, analytic – constraints of rationality. Nothing ethically meaningful is lost by conceding that the will is not free to violate these sorts of analytical constraints.

11 Aside from the entirely unrelated context of mathematics.

https://doi.org/10.1017/S1369415400000947 Published online by Cambridge University Press