A Kantian Theory of Evil

Is there any interesting sense in which we can speak of an act as 'evil', in contrast to simply "morally bad' or "immoral"? In ordinary language, we typically judge actions as evil that somehow differ significantly, in terms of degree or intensity, from commonplace wrongdoing. That is, what we regard as evil are just those actions that, to some greater extent, more seriously offend our deeply-held moral sentiments or that produce much more harmful consequences. We feel that brutal acts of murder, rape, torture, and mass genocide somehow go beyond immorality. We want to call them "evil".

If taken to an extreme, however, this view simply reduces the difference between evil and immoral acts to a mere quantitative analysis. At worst, it leads to a wholly uninteresting account of evil as just those actions we tend to regard as "really bad". In this paper, I want to sketch out a distinctively Kantian theory of evil that instead defends a fundamental *qualitative* difference between evil and more ordinary immoral actions, locating the main distinction in terms of *the structure of the agent's will itself.* Broadly understood, the present Kantian strategy endorses an account of "evil as dehumanization" in which, in a manner to be discussed below, a "material" (as opposed to purely "formal") difference exists between the respective maxims of the immoral and the evil agent—where, unlike typical cases of Kantian immorality, direct violation of another person's humanity *qua* human somehow comprises a necessary part of the "material object" of an evil agent's will.

The overall argument for this view proceeds in three stages. First, in §I, I outline Kant's own historical account of both moral and immoral action. Second, in §§II–IV, I suggest a novel strategy for modifying Kant's ethical theory in order to allow for a distinctive conception of evil, as opposed to simply immoral, action, and then identify at least four paradigm cases of "dehumanization" that appeal to certain basic moral intuitions about the

"A Kantian Theory of Evil" by Ernesto V. Garcia,
The Monist, vol. 85, no. 2, pp. 194–209.

nature of evil action. Third, and lastly, in §V, I offer concluding remarks about how this revised Kantian picture incorporates, in interesting ways, some historically influential views about evil.

I

From the outset, however, one may worry that the present undertaking is deeply misguided. For while recognition of any interesting difference between evil in contrast to immoral acts is quite seldom found in modern moral philosophy, Kant's ethical theory itself might be fairly interpreted as in fact opposed to such a distinction.

Recall that in the 1788 *Critique of Practical Reason*, Kant famously distinguishes between two senses of 'good' and their respective negative counterparts: (a) 'the good' [das Gute] in contrast to 'the evil' [das Böse]; and (b) 'the good' [das Wohl] in contrast to 'the bad' [das Übel or das Weh] [Ak. 5: 59–61/52–53].[1] On this Kantian model, to call something "bad" in the sense of (b) means to take it as somehow harmful or disagreeable to one's general "well-being," that is, *a state of affairs* resulting from wholly fortuitous unfortunate circumstances—as when, reflecting the original German terms used by Kant, we speak of a particular individual's "weal and woe." By contrast, to call something "evil" in the sense of (a) means to take *some action* as morally wrong, resulting not from natural contingency, but rather, from a direct act of the agent's will, when she wrongly subordinates the principle of morality to the principle of self-love.[2] Granted this distinction, what Kant himself calls "evil" should apparently be just subsumed under the wider rubric of "immoral acts" in general. Thus, one might suppose here, in a strict Kantian framework, no unique conception of "evil"—in contrast to mere "immoral" actions—is even permitted.

To demonstrate that a Kantian account of evil is plausible, then, we need to show how Kant's own characterization of evil does not exhaust all the possible ways in action may be morally deviant. I want to outline below Kant's changing historical views about both (a) moral and (b) immoral action. In doing so, I try to demonstrate how, particularly by drawing upon his later ethical views, we can obtain helpful conceptual resources to account for the idiosyncratic nature of evil action. I propose here that evil be regarded as a kind of morally corrupt *hybrid species of action*, one borrowing

elements from Kant's description of both moral and immoral actions, where the "object" of our will is a concrete state of affairs that just consists in direct violation of the humanity of another person itself: a conceptual possibility, I argue, Kant himself left unexplored.

A. Moral Action

In his 1785 *Groundwork for the Metaphysics of Morals* [Groundwork], Kant notoriously defends a very rigid "formalistic" account of maxims, or "subjective principles of action," that we act upon when obeying the moral law. As Kant writes:

> . . . an action from duty has its moral worth *not in the purpose* to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of action but merely upon the *principle of willing* in accordance with which the action is done. [Ak. 4: 399–400/13, emphasis in original]

In assessing the moral worth of an action, all that matters is the "form" of the maxim, and not any "matter" produced by our actions [Ak. 4: 400/13]. Nevertheless, for Kant, despite denying the moral relevance of a specific end we somehow bring into existence through our actions, some "end" is still always necessarily involved in morality. Such ends are termed "objective ends," comprised of the "humanity" of all rational beings regarded "as an end in itself" [Ak. 4: 429/37]. Kant later describes "humanity" as the "matter," that is, the sole necessary "end" of the principle of morality [Ak. 4: 436/43–44]. Our moral duty, however, is not to bring such "objective ends" into existence. Rather, "humanity" in each person is an "existent end" that serves as a formal "limiting condition" of our actions, requiring us both to refrain from certain deeds and to act harmoniously with it.

But is this mere limiting condition of humanity the only object of moral willing? In his later writings, Kant importantly qualifies this formalistic picture. Speaking again of "ends" in the 1793 *Religion Within the Boundaries of Mere Reason* [Religion], Kant writes there:

> An end is *always* the object of *inclination*, that is, of an immediate desire to possess a thing by means of one's action, just as a *law* (which commands practically) is the object of *respect*. An objective end (i.e., an end which we ought o have) is one which is assigned to us as such by reason alone. The end

> that contains the inescapable, and at the same time sufficient, condition of all other ends is the ultimate end. One's own happiness is the subjective ultimate end of rational beings belonging to the world. . . . But that every human being ought to make the highest possible *good* in the world his own *ultimate end* is a synthetic practical proposition *a priori*, that is, an objective-practical proposition given through pure reason. . . . That is, the proposition, "Make the highest good possible in this world your own ultimate end," is a synthetic proposition *a priori* which is introduced by the moral law itself, and yet through it practical reason reaches beyond the law. *And this is possible because the moral law is taken with reference to the characteristic, natural to the human being, of having to consider every action, besides the law, also as an end* [Ak. 6: 7/35, final emphasis added].

In Groundwork, Kant had employed a similar premise—that "what serves the will as the objective ground of its self-determination is an end"—to establish the fact that we must regard humanity as the necessary formal end of moral action [Ak. 4: 427–28/36]. Now in this Religion passage, Kant again enlists the fact that we, as rational agents, must have some end in view in order to ground the demand for a more "concrete" end, an actual state of affairs that we should aim at in moral activity. Morality requires that we be concerned about an "ultimate end"—the "highest good"—as an object of our will which, unlike humanity, does not act as a mere "limiting condition" for action, but instead, as a concrete, albeit quite temporally distant state of affairs that we must strive to bring to existence.

Finally, in his 1797 *Metaphysics of Morals* [MM], Kant introduces a wholly novel idea into his ethical theory, what he calls "ends which are at the same time duties" [Ak. 6: 382–4/147–8]. Now, instead of a future object of moral willing, Kant identifies here an immediate state of affairs that we ought to bring about in moral action: namely, the two "obligatory ends" of our own moral perfection and the morally permitted happiness of others. What is most relevant for us here is to recognize that within all these doctrinal shifts, Kant ultimately identifies the "matter" of moral action with two different types of "objects": (1) humanity viewed as an already "existent end" we must not act against; and (2) some concrete state of affairs we must bring about in moral action, including (a) the "highest good" as a future realization of perfect justice in which human happiness exists in exact proportion to virtue and (b) the two immediate "obligatory ends" in the *Metaphysics of Morals* of our own moral perfection and the permitted happiness of others.

B. Immoral Action

The foregoing account of moral action contrasts sharply with Kant's views about the nature of immoral action. In this section, I will restrict most of my attention to the interesting but almost entirely neglected historical development in Kant's doctrinal views about immoral action from the 1785 Groundwork to the 1788 CPrR, primarily because Kant's analysis of the basic structure of immoral action, I believe, remains essentially unchanged from the CPrR to the 1793 Religion and 1797 MM.[3]

It is easily overlooked that Kant's primary goal in the 1785 Groundwork is, perhaps surprisingly, to differentiate moral actions not so much from *immoral* ones, but rather, from *heteronomous* actions in general, where these may include maxims containing both (a) non-moral hypothetical imperatives and (b) immoral hypothetical imperatives.[4] Thus, Kant discusses moral action not only in terms of its differences with immoral actions like lying, suicide, and not helping others. Equally importantly, Kant contrasts moral action with merely sympathetic or prudential action (Section I) and with actions obeyed out of conformity to the will of God, moral feeling, or some ideal of perfection (Section II). Presumably, all of these latter actions may, and indeed often typically do, produce results that have mere "legality"—that is, they conform to what morality requires. For Kant, what is entirely absent is their possession of any intrinsic moral worth, as well as the guarantee that such actions shall never go morally astray.

What distinguishes moral action from such heteronomous activity in general? When we act heteronomously, we always obey hypothetical, as opposed to categorical, imperatives. In discussing heteronomous action in general, Kant writes:

> Wherever an object of the will has to be laid down as the basis for prescribing the basis for prescribing the rule that determines the will, there the rule is none other than heteronomy; the imperative is conditional, namely, *if* or *because* one wills this object, one ought to act in such and such a way; hence it can never command morally, that is, categorically. Whether the object determines the will by means of inclination, as in the principle of one's own happiness, or by means of reason directed to objects of our possible volition, as in the principle of perfection, the will never determines itself *immediately*, just by the representation of the action, but only by means of an incentive that the anticipated effect of the action has upon the will. . . . [Ak. 4: 444/50]

The "ought" here depends entirely upon the fact that we conditionally hold some "object" as our end. Furthermore, the specific "object" or "material end" of heteronomous activity is never provided to us by reason; rather, the end involved is only realization of some concrete state of affairs determined wholly *a posteriori* by what Kant terms natural "incentives" [Ak. 4: 427–8/36]. This stands in stark contrast to morality, where not only is the "end" of moral action—humanity—specified for us *a priori* by reason itself, but also, at least in the Groundwork, specific mention of some concrete state of affairs is entirely excluded from Kant's analysis of moral action.[5]

Arguably, it is only in Kant's 1788 CPrR that we find a detailed picture of the nature of *immoral*, in contrast to merely *heteronomous*, action. It is true that in the Groundwork, Kant describes immoral actions as those that involve maxims we cannot universally will without contradiction [Ak. 4: 421–5/31–33]. This analysis only tells us, however, what immoral actions are *not like*. What is lacking is any substantive positive construal of the nature of immortal action, an issue now fully addressed by the CPrR.

In the CPrR, Kant asserts that only two principles govern our action: the principle of self-love and the principle of morality. Kant officially characterizes the principle of morality in the CPrR thus: "So act that the maxim of your will could always hold at the same time as a principle in the giving of universal law [Ak. 5: 30/28]." About the principle of self-love, Kant writes instead:

> All material practical principles are, without exception, of one and the same kind and come under the general principle of self-love or one's own happiness.
>
> Pleasure arising from the representation of the existence of a thing, insofar as it is to be a determining ground of desire for this thing, is based on the *receptivity* of the subject, since it *depends upon the existence of an object.* . . . It is, then, practical only insofar as the feeling of agreeableness that the subject expects from the reality of an object determines the faculty of desire. [Ak. 5: 22/19, emphasis in original]

In the Groundwork, Kant had distinguished between "material" practical principles, which involve relative ends, and "formal" practical principles, which involve the objective ends of humanity [Ak. 4: 427–28/36]. In the CPrR, Kant now identifies all possible material practical principles under a single category, *viz.*, as specific instances of the principle of self-love.

But, if all actions besides moral ones fall under self-love, then a basic worry arises: can we ever clearly distinguish between non-moral and immoral actions? Kant himself characterizes the main difference between them in terms of the "formal" ordering of our will. In non-moral maxims, we obey the principle of self-love, where no relevant moral concerns conflict with our maxims; in immoral maxims, however, we set up the principle of self-love in direct opposition to the principle of morality, and wrongly subordinate demands of morality to our own happiness. As Kant describes this relationship in the section of the CPrR entitled "Incentives of Pure Practical Reason":

> This propensity to make oneself as having subjective determining grounds of choice into the objective determining ground of the will in general can be called *self-love*; and if self-love makes itself lawgiving and the unconditional practical principle, it can be called *self-conceit*. Now the moral law, which alone is truly objective (namely objective in every respect), excludes altogether the influence of self-love on the supreme practical principle and infringes without end upon self-conceit, which prescribes as laws the subjective conditions of self-love. [Ak. 5: 74/64, emphasis in original]

This is related to what Kant earlier affirms in this same section, when he writes:

> Pure practical reason merely *infringes upon* self-love, inasmuch as it only restricts it, as natural and active in us even prior to the law, to the condition of agreement with this law, and then it is called rational self-love. But it *strikes down* self-conceit, since all claims to esteem for oneself which precede conformity to the moral law are null and void. . . . [Ak. 5: 73/63, emphasis in original]

Immoral actions, then, are to be explained not only, as in the earlier Groundwork account, in terms of their inability to be universally willed without contradiction. Additionally, immoral actions now consist in a wrongful attempt by self-conceit to establish a relative maxim as if it were a genuine "categorical imperative," as if it were an "unconditional practical principle" endowed with lawgiving force somehow superior to morality itself.

II

Even granted our admittedly cursory sketch above of Kant's views about both moral and immoral actions, we are now equipped to see what a Kantian theory of evil might be like, in contrast to ordinary immoral action. For Kant, as we have observed, immoral actions involve at least

two essential elements: (1) the end involved in immoral action is always just some material state of affairs; and (2) the basic difference between moral and immoral actions is located in the formal structure of the will, where, in immorality, the principle of self-love comes into conflict with, and positively subordinates, the principle of morality—in contrast to mere non-moral action where no actual opposition with morality need take place.

In what follows, I want to defend a Kantian account of "evil as dehumanization," interpreting evil actions as qualitatively distinct from more ordinary immoral ones insofar as they fundamentally differ in a *material* sense: that is, insofar as they have an entirely different sort of "object" involved in their maxims. In immoral action, our object is some specific state of affairs, where the humanity of another person figures only *indirectly* into our actions as a useful *incidental means* towards the realization of this independent aim. Consider, for example, that, for the sake of happiness, my end is to obtain ready money in order to improve my material circumstances. In this case, when I make a false promise, I treat the humanity of another person as only a means: what Kant describes as a mere "effective cause" in a broader causal nexus, initiated by some natural incentive, and resulting in some concrete state of affairs that ultimately fulfills my heteronomous desire.

I propose that in cases of "evil as dehumanization" the humanity of another person is not just incidentally involved in our aims as some necessary means; instead, in evil action, humanity itself comprises a *constitutive* element of our end. More precisely, the object of our willing *just consists* in the mistreatment of the humanity of another person *qua* human—rather than conceiving of their humanity as just a means, as simply a useful "effective cause" for the realization of some *further* state of affairs. In evil actions, we seek out to directly violate the humanity of another person itself.

Notably, Kant himself never envisioned such a possibility. Indeed, on this view, evil actions represent a sort of morally corrupt hybrid between Kantian moral and immoral action: where they possess a *formal* similarity with immoral actions, and a *material* similarity with moral ones. Thus, like Kant's mature views about moral action, evil action would include both a concrete state of affairs and humanity itself as part of its "material end," although the state of affairs involved here simply consists in *the violation of humanity itself.*

And, like Kant's account of immoral action, evil actions always entail a formal subordination of the principle of morality to the principle

of self-love. There nonetheless remains a crucial distinction between evil and ordinary immoral action, because, as noted above, a fundamental material difference exists between them. In a sense, then, evil can be seen as a kind of perverse mimicry of moral action, where humanity is indeed treated as an "end in itself," but now for immoral acts of the will.

III

What are some possible candidates for this conception of evil? I want to identify here at least four fundamental types of evil that capture many of our basic intuitions about genuinely evil action, although obviously not exhausting all possibilities. The attempt to somehow directly violate the humanity of another person is a shared essential feature of each evil action discussed below. Differences between them amount to taking up various blameworthy modes of evaluation: (a) indifference to the destruction of the humanity of another person; (b) denial of the humanity of another person; (c) servility, or depreciation of one's own humanity; and (d) delight in suffering of the humanity of another person. I take up each particular case in turn below.

For the first kind of evil, we might consider the example of Adolf Eichmann. In this act of sending thousands of Jews to their deaths, Eichmann notoriously defended his actions by stating that he was simply obeying his orders. As Hannah Arendt reports in *Eichmann in Jerusalem*[6], Eichmann remarkably believed that by acting this way, he was being a good Kantian. In Eichmann's own eyes, Arendt writes, "He did his *duty*, as he told the police, and the court over and over again; he not only obeyed *orders*, he also obeyed the *law* [135]." What most disturbs us about this case? It seems to be not only Eichmann's purported justification of his actions in terms of obeying the law. It is the fact that, in the activity of dutifully obeying the orders of his superiors, Eichmann could be so indifferent to the humanity of his own victims. He somehow remained unaffected by the absolute value of the humanity of those persons whom he destroyed—wholly subordinating, in Kantian terms, the "matter" of the law to its mere, albeit radically distorted, universalistic "form."

For the second kind of evil, what perhaps come most readily to mind are cases of racist behavior. What seems most truly evil about racist action is not so much any particular acts of taunting, racial slurs, or displays of

disrespect, but instead, we might feel, the pernicious underlying attitude: that the racist sincerely believes members of a different race somehow count as less than human. Here importantly, unlike the first case, the racist is *not* indifferent to the value of humanity. Indeed, it is clear that she in fact *positively* values humanity, presumably taking pride in the intrinsic worth of her own humanity and of all others in her particular race. What she simply denies, however, is the thesis that members of some particular race are in fact human, and therefore deserving of respect or esteem. In such cases, we want to believe that this attitude cannot be excused as merely a theoretical mistake. If the racist were to observe with unprejudiced eyes the actual life of a person of a different race, we might feel that she could not honestly deny their humanity or persist in her belief that being a member of a certain race makes that person less human in any relevant way.

That is, one might believe that racism, if still embraced in spite of deep personal contact with people of different races, most likely involves some degree of self-deception in order to sustain this denial of humanity.[7] In fact, one might see the case of indifference to humanity as involving a similar sort of self-deceptive stance. For, insofar as we are aware that we are involved in the destruction of the humanity of another person whose value we recognize, we feel that we cannot remain indifferent to them, that we cannot honestly try to defend our actions by asserting that we were just obeying orders.[8]

Interestingly, the third and fourth kinds of evil can be viewed as negative mirror images of each other. While servility involves depreciating or attaching a kind of negative value to one's own humanity, while attaching a positive value to the humanity of our superior (and to their deriving pleasure from the fulfillment of their particular ends), in cases of delighting in the suffering of others, such as cruel cases of torture, we now attach a positive value to our own humanity (and to our "ends" as worthy of fulfillment) and take pleasure in depreciating or attaching a kind of negative value to the humanity of another person. Kant's own famous critique of servility in the MM helpfully elaborates the sort of dehumanization involved in such acts:

> In the system of nature, a human being (homo phenomenon) is a being of slight importance and shares with the rest of the animals, as offspring of the earth, an ordinary value. . . . But a human being regarded as a *person*, that is,

> as the subject of a morally practical reason, is exalted above any price; for as a person (homo noumenon) he is not to be valued merely as a means to the ends of others or even to his own ends, but as an end in himself, that is, he possesses a dignity (an absolute inner worth) by which he exacts *respect* for himself from all other rational beings in the world. He can measure himself with every other being of this kind and value himself on a footing of equality with them.
>
> Humanity in this person is the object of the respect which he can demand from every other human being, but which he must also not forfeit . . . he should not disavow the moral self-esteem of such a being [namely, a rational human being], that is, he should pursue his end, which is in itself a duty, not abjectly, not in a servile spirit (animo servili) as if he were seeking a favor, not disavowing his dignity, but always with consciousness of his sublime moral predisposition. . . . [Ak. 6: 435/186–7]

In cases of servility, we disavow *our own dignity*, and privilege the pleasure of *another* person whose favor we are seeking above the value of our own humanity. In cases of torture, on the other hand, it seems we instead deny the human dignity of *another* person, and privilege *our own pleasure* above the value of their humanity. In sum, the acts of valuing involved in both cases seem to be identical, although radically inverted: we inconsistently attach both a negative and positive value to humanity (whether our own humanity or the humanity of another person) as well as privilege the realization of some natural state of affairs (either the pleasure of that person whose favor we are seeking or our own pleasure) over the intrinsic value of humanity itself.[9]

IV

In discussing the cases of evil above, the observant reader will have noticed that I have just been modeling these examples after Kant's own typology of possible moral evaluation of actions outlined in the 1793 Religion. As Kant writes in a footnote:

> If the good = a, the opposite contradicting it is the not-good. Now, this not-good is the consequence either of the mere lack of a ground of good, = 0, or of a positive ground antagonistic to the good, = − a; in this latter case, the not-good can also be called positive evil. [6:23/48]

Given Kant's own categorization, we can offer the following analysis about the different modes of valuation involved in each example of evil action: (1) in cases of indifference to humanity, we give humanity no

value (0); (2) in cases of denial of humanity, we give humanity a positive value (+), but simply (and perhaps dishonestly) deny the thesis that another person of, say, a different race, is in fact human; (3) in cases of servility, we attach a negative value (−) to our own humanity, and a positive value (+) to the humanity of another person (and their particular ends), whereas (4) in cases of, say, torture, we attach a negative value (−) the humanity of another person, while attaching a positive value to our own humanity (and its particular ends), where in both instances, we wrongly privilege some entirely natural state of affairs—namely, the realization of pleasure, whether our own or another person's—over the intrinsic moral value of humanity itself.

Identifying these cases as paradigmatic examples of evil action arguably captures, while of course not exhausting, many of our ordinary intuitions about when certain actions are evil. Clearly, granted our present typology, the examples above leave out at least three final modes of evaluation: cases of attaching a positive value (+) to the humanity of *both* persons; cases of attaching a negative value (−) to *both* persons; and cases where we are not engaged with attaching any value whatsoever.

Obviously, all instances of attaching a positive value to humanity, if properly undertaken, just amount to genuine moral action. The last two cases are perhaps more controversial. For the first kind of evaluation, consider the case where someone commits a crime so horrifying that we are inclined to attribute to that agent a loss of all moral sentiment, and for the second, consider the case of a misanthrope who hates humanity in general.

Kant himself offers an interesting analysis of the first kind of case when discussing what he refers to as "great crimes" in the MM, writing:

> [I]t is not only unnecessary but even improper to ask whether great *crimes* might not require more strength of soul than great virtues . . . great crimes are paroxysms, the sight of which makes one whose soul is healthy shudder. The question would therefore come to something like this: whether a human being in a fit of madness could have more physical strength than when he is sane. This one can admit without attributing more strength of soul to him, if by soul is meant the vital principle of man in the free use of his powers; for, since the basis of great crimes is merely the force of inclinations that weaken reason, which proves no strength of soul, the above question would be tantamount to whether someone could show more strength during an attack of sickness than when he is healthy. [Ak. 6: 384/148–9]

Here, a certain action seems to be so brutal and horrifying that it goes beyond the domain of what humans are capable of, and hence we believe

that the agent has somehow lost all moral sentiment whatsoever: that is, she appears to have stepped *outside the realm of human valuation altogether*. Thus, in such cases of true loss of moral sentiment, we do not ordinarily interpret her as occupied with the typical business of assigning moral value at all—whether negative, positive, or none—either for her own humanity or for the humanity of another person.

In the case of the misanthrope, on the other hand, we might instead feel that she not only indeed acknowledges the fact that the humanity of other people possesses a negative value, but, if she is fully consistent in her beliefs, she must recognize that she, too, as human, ought likewise to be regarded as an object of hatred. That is, in this second case, we think the misanthrope most likely believes that humanity in general has a negative value, both when found in other persons and in herself. But, if this is so, then the misanthrope more accurately seems to elicit from us not so much moral outrage as sheer pity. We are rather inclined to feel deeply sorry for, rather than angry at, a misanthrope who cannot find any genuine positive value in humanity itself, either in the humanity of other persons or presumably even in her own.[10]

V

In this paper, I have tried to outline an account of a distinctive Kantian theory of evil understood as dehumanization, as well as to provide a detailed analysis of a few different types of evil action. I close here with two observations. First, in terms of Kant scholarship, insofar as one agrees with this analysis, the present discussion lends support to the widespread intuition that Kant's "Formula of Humanity" [FH] constitutes a philosophically better version of the Categorical Imperative than the more traditionally influential "Formula of Universal Law" [FUL].[11] For it seems difficult to see how FUL would capture the distinctive way in which, when we act in a truly evil way, we somehow directly seek to dehumanize the humanity of another person. A unique analysis of evil actions, in contrast to immoral ones, seems as if it would be lost in a merely formal analysis of whether or not we can universally will some specific maxim.

Second, and of broader ethical interest, I think this account of evil captures both historically influential as well as many modern sentiments we have about the notion of evil. Parallel to more traditional views of evil,

either in terms of a philosophical account of evil as "privation" of being, or else in terms of a theological account of evil as a "transgression" against God's will, the Kantian theory of evil defended here can be seen as both capturing, as well as radically transforming, intuitions underlying both ideas. On the present account, evil is clearly regarded as somehow a fundamental privation of our own humanity. And it can be viewed as transgressing, not against divine command, but now instead against the absolute value of the humanity of another person as legislated *a priori* by human reason itself, thereby preserving a notion of evil within a wholly secular context.

Finally, insofar as we think that Kant's idea of dignity captures well many of our present-day moral intuitions, I have tried to demonstrate here how evil action can be plausibly construed as a direct assault upon this fundamental modern ideal of the inherent dignity of all persons. In his book *Individualism*, Steven Lukes notes that "recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family," as stated by the 1948 U.N. Universal Declaration of Human Rights, finds:

> its most impressive and systematic expression in the writings of Immanuel Kant, who asserted that 'man, and in general every rational being, exists as an end in himself, not merely as a means for arbitrary use by this or that will: he must in all his actions, whether they are directed to himself or to other rational beings, always be viewed at the same time as an end'.[12]

The present Kantian theory of evil highlights the importance of the value of humanity in moral evaluations. Evil understood as dehumanization fundamentally distorts and undermines, in the case of both the victim and the evil agent herself, the intrinsic absolute value of our own shared humanity.[13]

Ernesto V. Garcia

Columbia University

NOTES

1. The following Kant texts are cited: *Groundwork for the Metaphysics of Morals*, translated by Mary Gregor (Cambridge: Cambridge University Press, 1997); *Critique of Practical Reason*, translated by Mary Gregor (Cambridge: Cambridge University Press,

1997); *Religion within the Boundaries of Mere Reason*, translated by Allen Wood and George di Giovanni (Cambridge: Cambridge University Press, 1998); and *The Metaphysics of Morals*, translated by Mary Gregor (Cambridge: Cambridge University Press, 19960.

2. I discuss the fundamental Groundwork distinction between states of affairs in contrast to actions in §I, Parts A and B, below.

3. For present purposes, Kant's idea of radical evil, as well as the different classes of evil action in the 1793 Religion in terms of frailty, impurity, and perversity, are mostly irrelevant. (These notions deal more with the idea of the moral character with respect to one's life as a whole, rather than a more focused analysis of the basic structure of the will involved in any particular immoral action.) About Kant's interesting discussion of the nature of "great crimes' in the 1797 MM, however, see below, §IV.

4. The present interpretation makes sense in light of the stated aim of the preface of the Groundwork, where Kant writes that "the present groundwork is, however, nothing more than the search for and establishment of *the supreme principle* of morality. . . . " [Ak. 4: 392/5]. This presumably includes not only seeing how morality differs from immoral actions, but also, and perhaps more significantly, how it differs from spurious moral principles that prescribe actions which may conform with morality, but lack any genuine moral worth.

5. It should be noted that in his later ethical writings, Kant sees concrete consequences resulting from our actions figuring into morality, not in terms of specifying what we ought to do, or what makes an action morally good, but rather as simply taking into account that when we act, granted the structure of the human will itself, we always must have some end in view.

6. Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Viking Press, 1964).

7. If the racist could only, in a theoretical way, see a person of a different race as akin to, say, a rock, we would perhaps find her perceptions truly bizarre. It is because we think that the racist cannot think of another human being as just a rock, but instead, because we believe a person of a different race displays overt marks of humanity that the racist herself can potentially recognize, that we believe the racist is engaged, in a kind of self-deceptive way, in the deliberate suppression or active denial of any evidence that goes contrary to her own deeply-held convictions.

8. It is important that, historically, Eichmann was not like our case of the racist described above. He did not maintain the belief that Jewish persons were not human, and therefore possessed no intrinsic value, but in fact, early on, as Arendt tells us, actually tried at times to rescue certain groups of Jews, though in the end, he altogether abandoned all such efforts in favor of strict compliance with his orders.

9. Insofar as we think it is plausible to attribute intrinsic value to things other than humanity, the present analysis captures equally well intuitions about cases of evil action involved, say, when somebody delights, for example, in torturing a dog, or in defacing a picturesque instance of natural beauty. Kant famously calls duties to animals and to the environment as indirect duties, where we somehow are failing properly to respect our own humanity. On this view, they can be direct duties, if we recognize that such things possess intrinsic value like ourselves.

Kant, of course, would deny this possibility, because he believed that non-rational beings or states of affairs can only have relative value—see, for example, Groundwork, Ak. 4: 428/37.

10. It is perhaps instructive to compare these two cases—(a) a misanthrope who assigns negative value to all persons and (b) someone who has left the business of assigning value altogether (and with that, perhaps his very humanity)—with a final mode of evaluation—(c) a person who assigns zero value to all persons. This final option seems to embody the traditional "amoralist," who, as Joseph Raz describes him, "denies that all persons are valuable in themselves" (370). See his article "The Amoralist," pp. 369–98, in *Ethics and Practical Reason*, ed. by Garrett Cullity and Borys Gant (Oxford: Clarendon Press, 1997).

11. For cases of some commentators who philosophically privilege the FH over the FUL, see, for example, Christine Korsgaard, "The Right to Lie: Kant on Dealing with Evil," pp. 133–58, in *Creating the Kingdom of Ends*, (Cambridge: Cambridge University Press, 1996); Thomas Pogge, "The Categorical Imperative," pp. 172–93, in *Grundlegung der Metaphysik der Sitten: Ein kooperativer Kommentar*, edited by Ottfried Höffe (Frankfurt am Main: Vittorio Klosterman, 1989); and Allen Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999).

12. Steven Lukes, *Individualism* (Oxford: Clarendon Press, 1980), p. 49.

13. I would like to thank Thomas Pogge, Pat Kitcher, Wayne Proudfoot, Sirine Shebaya, and John Brunero for their many helpful comments and criticisms. Most important, I am grateful to Arik Ben-Avi, whose innumerable discussions have consistently made my ideas clearer and better, and to Amy Meselshon, with whom I share many of the ideas presented here, and without whom (though we differ in the basic analysis of the issues involved) I could not have arrived at my own formulation of these issues.