May 2021

# Beyond Depraved: Villainy and Self-Deception in Kant's Taxonomy of Evil

Kevin Alexander Korczyk
*University of Wisconsin-Milwaukee*

# BEYOND DEPRAVED: VILLAINY AND SELF-DECEPTION IN KANT'S TAXONOMY OF EVIL

by

Kevin A. Korczyk

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

in Philosophy

at

The University of Wisconsin-Milwaukee

May 2021

ABSTRACT

BEYOND DEPRAVED: VILLAINY AND SELF-DECEPTION IN KANT'S TAXONOMY OF EVIL

by

Kevin A. Korczyk

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Nataliya Palatnik

Kant's account of evil has often been criticized for being overly restrictive in that it seems unable to account for profoundly immoral acts such as those committed by the Nazis. In response, most defenders of Kant have attempted to gerrymander his original categories of evil such that they become expansive enough to account for these cases. In this paper, I argue that such defenses fail because they rule out the possibility of immoral acts committed *intentionally* and in *full knowledge* of their immorality. However, I also show that there is room in Kant's ethics for an additional category of evil that *can* account for such cases. I term this new category "villainy," and then point to a variety of self-deception that Kant ignores in order to help explain *how* villainous agents are able to intentionally do what they know to be wrong within a Kantian framework. Lastly, I defend the compatibility of "villainy" with Kant's ethics writ large by responding to four objections. First, that villainy violates Kant's commitment to the "guise of the good" principle; second, that villainy should only be possible for diabolical beings; third, that our lack of insight into agential intention forces us to remain skeptical about the existence of villainous agents; and fourth, that we risk "aestheticizing" evil if we admit that human beings can be villainous agents.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| *Gr* | *Groundwork of the Metaphysics of Morals* (In *Practical Philosophy*, Tr. Mary J. Gregor, Cambridge University Press, 2016) |
| *KprV* | *Critique of Practical Reason* (In *Practical Philosophy*, Tr. Mary J. Gregor, Cambridge University Press, 2016) |
| *RGV* | *Religion Within the Boundaries of Mere Reason* (In *Religion and Rational Theology*, Tr. Allen W. Wood and Geroge Di Giovanni, Cambridge University Press, 2005) |
| *MdS* | *The Metaphysics of Morals* (In *Practical Philosophy*, Tr. Mary J. Gregor, Cambridge University Press, 2016) |

## I. Introduction

Kant has frequently been accused of providing an account of human evil in *Religion Within the Boundaries of Mere Reason* that is too narrow to explain extreme acts of immorality. John Silber, for example, has argued that Kant's division of evil into the graded categories of frailty, impurity, and depravity fails to explain the apparent ability of human beings to contravene the moral law simply because we desire to do what is wrong (Silber 2012: 334). When levying this and similar criticisms against Kant, it has become standard practice to cite a litany of historical atrocities in evidence of the incredible depths of immorality to which human beings can sink.[1] Sadly, in this respect the 20th and 21st centuries provide no shortage of source material for us to draw upon. Take for example the current proliferation of mass violence in the United States, some instances of which reportedly lack any identifiable motive (e.g. the 2017 Las Vegas shooting).

More recently, important work has been done to respond to such challenges on Kant's behalf. Henry Allison has argued that Kant had good reason for denying that human action can be motivated by purely evil incentives, and that even the most heinous acts, when properly understood, can be accounted for within Kant's original categories of evil. One point that Allison appeals to is that, for Kant, the capacity to do evil for its own sake can only be a live option for "diabolical beings," i.e. beings who, unlike most human beings, do not recognize the moral law's authority, and who therefore are not morally responsible agents (Allison 1996: 176-181).[2]

[1] For examples of this phenomenon, see Arendt's *Origins of Totalitarianism* (1951), Silber's "Kant at Auschwitz" in *Kant's Ethics: The Good, Freedom, and The Will* (2012) and Card's *The Atrocity Paradigm: A Theory of Evil* (2002).

[2] I say "most human beings" here because it is possible that some human beings suffering from severe mental illnesses like sociopathy might turn out to constitute examples of diabolical beings. However, if this were the case, then sociopaths would fail to meet Kant's standards of human *personhood* (we will return to this point in §III.i). Sociopaths are therefore only tangentially relevant to this debate. (To be clear, Kant himself never explicitly mentions sociopaths. Yet, if human beings can be "diabolical" *at all*, then sociopaths seem like obvious candidates.)

While Allison is surely correct that we must be careful to distinguish between diabolical beings and merely evil human beings, I think that the initial criticisms of Kant's categories of evil remain compelling ones. Consider again the mass shooter. Assuming they are not sociopathic, it is difficult to imagine that such a person could *fail to recognize* the wrongness of murdering innocent strangers. Yet, mass shooters still choose to murder, and appear to intentionally disregard morality in the process. It is this kind of plausible scenario in which agents *knowingly* and *intentionally* do what is wrong that even the most charitable interpreters of Kant have difficulty explaining.

What I propose in this essay is a revision to Kant's taxonomy of evil in which a fourth category of evil is adjoined to Kant's initial three. In answer to criticisms such as Silber's, this additional category will involve a level of malevolence more profound than that of Kant's original categories, yet, in answer to Allison, it will not go so far as to entail the *complete* amorality of diabolical beings. I term this new category of evil *villainy*. Whereas Kant's diabolical being acts immorally from a sheer lack of respect for the moral law, villainous agents act immorally precisely *because* they rebel against the sense of moral humiliation that necessarily flows *from* the moral law. If we acknowledge the compatibility of villainy with Kant's practical framework, then we can coherently view some human agents as defying the moral law more directly than Kant thought possible, and can do so without denying that they are still morally responsible for their actions.

In the process of developing this new category of evil we will also disclose a new piece of moral psychology that Kant ignored, but that coheres with his overall project. That is, we will find that there is a possible form of *self-deception* about both the *source* and *value* of the moral law that, paradoxically, enables the *clear-sighted* perpetration of evil that characterizes villainous agents. That this form of self-deception seems possible should be of interest to anyone who thinks that self-deception plays an integral role in Kant's account of evil.[3]

[3] Many Kant scholars take this stance. Among those cited here are Allison (1996), Rukgaber (2015), and Papish (2018).

The paper will proceed as follows. In §II, I provide a brief outline of Kant's theory of action as it bears on the topic of immorality. Then, in §III, I interpret Kant's account of human evil as it is presented in *Religion Within the Boundaries of Mere Reason*. In §IV, I motivate the claim that Kant's own account cannot capture the full spectrum of human evil and then argue that this inadequacy can be compensated for by adopting villainy as a distinct category of evil. I then conclude §IV by addressing what I take to be the two strongest objections to my view: that villainous agents seem to violate what is commonly known as "the guise of the good" principle; and that villainous agents are constitute examples of diabolical beings. Lastly, in §V, I turn to two more minor objections that might be raised against my view: that the opaqueness of agential intentions gives us reason to doubt the existence of villainous agents; and that admitting of the possibility of villainy may lead to a dangerous "aestheticization" of evil. If all goes according to plan, the current incompleteness of Kant's account of evil will be revealed, but my proposed revision will give us reason not to take this as a decisive problem for Kant's overall view.

## II. Kant On Action and Immorality

The goal of this section is to show that Kant's views on human action commit him to the claim that all immorality involves *practical irrationality*. Since this interpretation of Kant is employed throughout the essay it is necessary to begin by explaining it in some depth.

Broadly speaking, human beings are rational agents. According to Kant, this means that we act on principles, some of which are *subjective* and hold only for us (i.e. *maxims*), and others of which are *objective* and hold for all rational beings (i.e. *practical laws*) (*KprV* 5:19). We can think of the objectivity of practical laws as consisting, in part, in their constituting the *standards of rationality* against which our maxims can be judged (Timmons 1994: 120). Yet, practical laws also serve another purpose. For Kant, the will itself *just is* the capacity of reason to guide action by means of principles (*Gr* 4:412).

Hence, if we are to think of human wills as *free*, which for Kant is necessary for moral responsibility, we must first identify a special *self-legislated law* that can govern free wills (*Gr* 4:449-4:450). This self-legislated law, Kant argues, is none other than the *categorical imperative*: the *moral law* (*Gr* 4:447, *KprV* 5:27).

Kant thinks of the adoption of a maxim, represented by an "I will…" statement, as the formation of an intention to act in a particular way. These maxims are always taken up by agents on the basis of motivational reasons, or what Kant calls *incentives*.[4] However, agents are not merely passively moved by incentives, but rather, they *freely* "incorporate" incentives into their maxims as the "subjective determining ground" of their will (*KprV* 5:72).[5] It is this freedom with regard to motivations that makes agents *responsible* for the incentives that they select for their maxims. Allison calls this view of Kant's "the incorporation thesis." (Allison 1990: 5-6, 39-40).[6]

The incorporation thesis is important for Kant because he thinks that incentives can be either *moral*, when the moral law itself is our motivation for adopting a maxim, or based in *self-love*, when our motivation is our own happiness or desires (*RGV* 6:36). Since we often allow both kinds of incentives to play a motivating role in our actions (e.g. when we do something both because it is moral *and* because it gives us pleasure), it becomes possible for us to *subordinate* the moral incentive to non-moral incentives in our maxims (*RGV* 6:36). That is, it is possible for us to take the moral

[4] Both Timmons and Allison agree that incentives can be profitably thought of as the motivational reasons for our actions (Timmons 1994: 117, Allison 1990: 40).

[5] The "*objective* determining ground" of our will, on the other hand, is in every case reason itself, viz. the practical laws of reason. That is, our will is always guided by the categorical and hypothetical imperatives (qua practical laws) even when our maxims fail to meet their standards. For this reason, it is sometimes said that they are "constitutive" of willing itself (Korsgaard 2013: 81).

[6] In the course of the discussion so far, we have identified two functions, or powers, of the will. Firstly, the will enables us to derive actions from the representation of principles, and secondly, the will allows us to freely select which incentives we incorporate into our maxims. The first power of the will is usually identified as that which Kant refers to as *Wille*, often translated as "will", while the second power is identified with what Kant refers to as *Willkür*, often translated as "the power of choice" (Beck 1960: 178, Allison 1990: 130).

incentive as motivating for us *only* in case we have incentives from self-love for the action in question. To put it pithily, subordinated incentives are those that only motivate us *conditionally*.

Here is the point: Kant tells us that whenever agents subordinate the moral incentive to non-moral incentives in their maxims, the resulting actions are *immoral* (*RGV* 6:36). This follows from the fact that doing so necessarily involves a breach in the standards of rationality as dictated by the practical laws of reason. Given that the categorical imperative (the moral law) is itself a practical law of *reason*, subordinating morality to self-love is *irrational* (Timmons 1994: 120). For Kant then, immorality is inherently bound up with *practical irrationality*.

If we understand immorality in terms of practical irrationality, as I think we should, then immorality will not consist *solely* in the inverted ordering of incentives outlined above. Rather, immorality will also consist in any *error in practical rationality* that *promotes* our subordinating the moral incentive. As we will see in the following section, the primary kind of rational error that Kant concerns himself with in his ethics is *self-deception*.

## III. Kant's Taxonomy of Evil

We are now in a position to elucidate Kant's account of evil and its categories. We will begin with a brief discussion of Kant's views on human nature before turning to the categories themselves.

### III.i Kant on Human Nature

According to Kant, human beings can possess either a good character or an evil character; it is impossible to straddle the fence. What determines an individual's character is her commitment to a very general maxim, which Kant calls a "supreme maxim." The supreme maxim is always one of two varieties: it is either a maxim in favor of universally prioritizing moral incentives over all others, or it

is not. In the former case an agent's character is good, and in the latter case it is evil (*RGV* 6:31).[7] More controversially, Kant also claims that *all* human beings have a tendency to opt for the evil variety of supreme maxim, and that we therefore can be said to possess "a propensity to radical evil" (*RGV* 6:32).[8]

Despite the fact that human beings incline toward evil, Kant claims that we also have three ineradicable predispositions that promote good action (RGV 6:26). The first two, the predispositions to animality and humanity, can be categorized under the umbrella of self-love. The predisposition to animality is the source of our incentives for preserving our own life and the life of our species, as well as of those incentives that fuel our social drives. When this predisposition is not tempered by a commitment to the moral law Kant claims it leads to vices such as gluttony, lust, and "wild lawlessness" with respect to others (RGV 6:26-27).[9] The predisposition to humanity, on the other hand, is the source of our incentives to seek our own happiness by comparing our condition with that of others. Kant thinks that this predisposition, when given excessive influence, produces the vices of jealousy and of joy in the misfortune of one's rivals (RGV 6:27). The third predisposition, the predisposition to personality, is the most important. It consists in the mere fact that we recognize the moral law's authority over us and possess the moral feeling of respect. Hence, the predisposition to personality consists in our awareness of the moral law, or as Kant would say,

---

[7] Kant thinks that the supreme maxim is something that we find active in us from the moment that we begin to deliberate about action, which makes it unlike any other maxim. Yet, it is also crucial to Kant that it be possible for us to *change* our supreme maxim through a moral "revolution" in our character (*RGV* 6:37, 6:47). If moral progress of this kind was not possible for us, then it would be impossible to ever *fully* obey the moral law's commands (since it demands moral perfection).

[8] There has been much debate as to how this universal claim about the human propensity to radical evil can be made consistent with Kant's acknowledgment of human freedom. Most interpreters agree that at least part of the answer lies in Kant's claims about the "unsocial sociability" of human beings (cf. Wood 1991; Allison 2002; and Allais 2017). That is, most human beings are led to immorality on account of a perversion of the spirit of *competition* that necessarily arises from the fact that we are self-interested individuals on the one hand, and members of a society on the other.

[9] Kant never elaborates on what "wild lawlessness" is supposed to mean. However, he does imply that it is a vice grafted onto our incentives for "community with other human beings" (*RGV* 6:26-6:27). A plausible reading then is that it is a vice involving criminal activities that exploit our social relations with others (such as theft, adultery, or violence).

our consciousness of the moral law as "the fact of reason" (K*prV* 5:31). It will be important for the ensuing discussion that, for Kant, it is only in virtue of the predisposition to personality that our status as morally responsible *persons* can be secured.

In sum, when we adopt an evil supreme maxim, our character is fundamentally tainted, and we therefore set ourselves up for immoral behavior. Yet, not all evil characters are created equal. Evil expresses itself in a variety of distinct shapes, or categories, some of which involve a greater degree of immorality than others. Kant specifies three: *frailty*, *impurity*, and *depravity*. Presumably, he intends these to exhaustively cover the types of immoral action that human beings are capable of committing, but as we shall see, his taxonomy of evil has difficulty capturing more drastic forms of immorality. Specifically, it seems incapable of explaining the existence of agents who knowingly and intentionally do what is wrong.

### III.ii The Categories of Evil

The first and least serious category of evil is frailty. An agent exhibits frailty when she possesses a genuine desire to do what the moral law commands in some particular case, but then irrationally fails to follow through on this desire in practice (*RGV* 6:29).[10] Consequently, she views her immorality as the result of an overpowering force within herself that she feels too weak to resist. According to Allison, an essential ingredient in frailty, and all three degrees of evil for that matter, is *self-deception* (Allison 1996: 178-179). Allison suggests that in order for her to view her actions as resulting from a "weakness" within herself, the frail agent must *deceive herself* about the fact that her

[10] To admit that the frail agent *desires* to do what the moral law commands with regard to one kind of action is not to say that she does not possess an evil character. The frail agent is simply someone who has adopted at least one very general moral maxim with the intent to live up to it, but who has also adopted an evil *supreme maxim*, which allows her to subordinate the moral incentive and thereby shirk her moral commitments. As Kant puts it, the frail agent is one in whom, "an evil heart can coexist with a will which in the abstract is good." (*RGV* 6:37).

actions result from her own free choice (Allison 1990: 157-161).[11] Kant himself lends some support to this interpretation in his *Metaphysics of Morals*:

> "if someone, from self-love, takes a wish for the deed because he has a really good end in mind, his *inner lie*, although it is contrary to his duty to himself, gets the name of *frailty*, as when a lover's wish to find only good qualities in his beloved blinds him to her obvious faults." (*MdS* 6:430; my emphasis).

What Kant says here is that the frail agent is someone who chooses to identify only with the aspect of herself that desires to do good and ignores the aspect of herself that is responsible for her immoral actions. In this way, the inner life of the frail agent is analogous to a "lover" who focuses only on her beloved's positive features while ignoring his flaws. As a result of her selective attention, the frail agent is blind to her moral inadequacy and is thus able to deceive herself into thinking that she simply possessed insufficient strength of will to act as morality commands.[12]

The second category of evil, impurity, characterizes an agent who conforms her actions to the moral law, and so views herself as good, but who in fact only acts morally because she has incentives from *self-love* that accidentally align with the moral law's commands (*RGV* 6:30). For example, an impure agent might adopt a maxim to refrain from stealing, but only because the threat of punishment is an incentive for her apart from the moral incentive. On the face of it then, the difference between frail and impure agents seems like a matter of "moral luck." That is, the impure

[11] Allison's interpretation of frailty and the role of self-deception in Kant's account of evil is not without controversy. See Timmons and Rukgaber for suggestions that self-deception is actually not necessary to account for the immorality of the frail agent (Timmons 2014: 124, Rukgaber 2015: 251).

[12] It is worth noting that the frail agent does in some sense *knowingly* do what is wrong, since she recognizes that her actions are immoral despite the fact that she attributes them to something in herself that is not herself. Yet, frailty as a category still cannot plausibly account for the sorts of extreme cases that primarily concern us here. Since the frail agent, by definition, truly desires to do what is good, it is clear that she is not an agent who knowingly *and intentionally* does what is wrong. That is, the frail agent does not initially *set out* to do what she knows to be wrong. We'll further discuss the insufficiency of Kant's categories to explain intentional immorality in §IV.

agent *just happens* to possess incentives from self-love for adopting maxims that are aligned with "the letter" of the moral law, while the frail agent does not.

I think a better way to distinguish between these two degrees of evil is to view impurity as involving a new variety of self-deception that was absent in frailty. Unlike the frail agent, the impure agent can be construed as deceiving herself about the fact that she frequently requires incentives from self-love to do what morality commands. One plausible way that an impure agent might accomplish this is by abstaining from introspectively investigating her own incentives to the degree that reason demands (Rukgaber 2015: 254-255). That is, the impure agent allows herself to believe that she acts from purely moral incentives by simply never bothering to reflect honestly on the nature of her incentives. This reflective negligence adds an additional level of practical irrationality, and hence immorality, to impurity that was absent in frailty.[13] While the frail agent at least recognizes that she acts from self-love and laments it, the impure agent rests content with an incomplete understanding of her own incentives.

All that remains is to examine Kant's final degree of evil: depravity. The depraved agent is someone who consistently subordinates morality to self-love in her maxims (*RGV* 6:30). A primary difference between depravity and the other two degrees of evil therefore lies in the *extent* to which the depraved agent disregards the moral incentive. For this reason, Kant writes that only depravity involves what can be called "deliberate guilt":

[13] As an aside, it is worth noting that Kant views human beings as possessing only a limited capacity to introspectively gain insight into our own incentives (*Gr* 4:407). However, as he indicates in *Religion Within the Boundaries of Mere Reason*, human beings are still morally responsible for, "screening incentives […] in accordance with the moral guide" (*RGV* 6:37). So, Kant seems to think that human beings, despite our limits, still commit a rational error whenever we do not attain at least the *imperfect* level of self-understanding that is within our reach.

> "as deliberate guilt (*dolus*), [depravity] is characterized by a certain *perfidy* on the part of the human heart […] in deceiving itself as regards its own good or evil disposition and, provided that its actions do not result in evil […], considering itself justified before the law." (*RGV* 6:38).

In this passage, Kant explicitly ties the moral corruption of depravity to self-deception. The depraved agent is someone who deceives herself as regards the *stringency*, or even the *content*, of morality by ignoring either the universal nature of its commands or by ignoring the absolute primacy of moral incentives. As Martin Sticker points out, for Kant, self-deception about the *stringency* of the moral law consists in convincing oneself that the moral law permits of exceptions in extraordinary cases where other considerations outweigh moral duty. On the other hand, self-deception about the *content* of the moral law consists in convincing oneself that the moral law *itself* contains concessions to self-love either in its criteria for what is moral or in its demands on the kinds of incentives that are appropriate to moral action. The distinction between self-deception about the moral law's stringency and self-deception about its content therefore consists in this: that those deceived about the moral law's stringency are *conditionally* committed to an accurate conception of the moral law, while those deceived about its content are *unconditionally* committed to a corrupt conception of the moral law (Sticker 2016: 90-91).[14]

One consequence of the depraved agent's self-deception is that it enables her to view herself as having a "good disposition," even when she has in fact merely externally aligned her actions with the moral law (*RGV* 6:38). Though both impure and depraved agents can take mere external alignment with the moral law as sufficient for the satisfaction of morality's demands, there is a crucial difference in the *way* in which they deceive themselves into taking up this stance. While the impure

[14] Of course, it is entirely possible that a depraved agent might be simultaneously self-deceived about both the moral law's stringency *and* its content. In this case, the agent would be *conditionally* committed to a *corrupt* conception of the moral law.

agent merely deceives herself about her incentives for action, the depraved agent deceives herself at the deeper level of *interpreting* the moral law itself.

A final key feature of depraved agents is that through their "lie to [themselves] in the interpretation of the moral law," they can even represent their immorality as *justified* so long as it does not result in obviously immoral actions (*RGV* 6:43). For example, a depraved agent might adopt an immoral maxim to pursue affluence even if it means doing dishonest business (thereby subordinating morality to self-love). Even so, it is possible that this agent might still always act in a way that is outwardly in accord with morality (e.g. if, through moral luck, she never has a good opportunity to do dishonest business). As a result, this agent may even take pride in her upright business ethics even though her maxim is corrupt.

In the above case, the depraved agent's justification for her immoral maxims depended on moral luck, but this needn't be the case. Another way that a depraved agent might justify her actions to herself is by engaging in *rationalization*. In the *Groundwork*, Kant writes that we as agents can develop,

> "a propensity to rationalize against those strict laws of duty and to cast doubt on their validity, or at least upon their purity and strictness, and, where possible, to make them better suited to our wishes and inclinations," (*Gr* 4:405).

Through rationalization, a depraved agent may make use of their self-deception about the moral law's stringency and/or content in order to represent even outright breaches of the moral law as morally justifiable.

We can illustrate this by slightly altering the example above. Imagine once again a depraved agent who has adopted an immoral maxim to pursue wealth at any cost. Now, suppose that this agent receives a once in a lifetime opportunity to sign a contract with a major corporation that will guarantee widespread distribution of the product that she manufactures. However, let us also

suppose that the contract has one caveat. The agent must guarantee that she can meet production demands without relying on the use of any cheap non-environmentally-friendly materials. Unfortunately, the agent recognizes that this stipulation would be impossible to satisfy since it would require a complete redesign of her product. Therefore, she signs the contract anyways and dishonestly portrays her product as environmentally friendly.

In this case, rather than accepting that what she has done is morally objectionable, the depraved agent might instead rationalize her action as a permissible exception to morality's commands. In virtue of a self-deceptive misinterpretation of the moral law's stringency for example, she might tell herself things like, "Sure, lying is usually wrong, but it's completely excusable when millions of dollars and my entire future livelihood is at stake." By engaging in rationalization of this sort, a depraved agent is able to improve the moral standing of her actions in her own eyes even when they constitute flagrant offenses against the moral law.

The above two cases illustrate what makes depravity a higher degree of evil than both frailty and impurity. While frail and impure agents are at least honest with themselves about what the moral law commands, the depraved agent deceives herself even in this regard so as to *justify* her immoral actions.

On the whole then, one can look at Kant's taxonomy of evil in two ways. First, one can see it as a hierarchical ranking of evil character types in terms of their involving progressively greater *irrationality*, and hence immorality (e.g. depravity's misinterpretation of the moral law adds a further level of irrationality to that found in impurity, and so on). Second, one can see Kant's categories as ranked in terms of increased degrees of *self-deception*.[15] So, if it turns out that human beings are

[15] Since self-deception is itself always an error in practical rationality, I think that these two interpretations ultimately amount to the same view. That is, if self-deception is itself a form of practical irrationality, then any hierarchical ranking of the categories in terms of greater or lesser degrees of self-deception will inevitably track a ranking in terms of greater or lesser degrees of irrationality.

capable of a more radical kind of immorality or self-deception than Kant admits of here, then a further category of evil will be required to explain this fact.

## IV. Villainy as a Distinctive Category of Evil

It is my contention that the account of evil presented in *Religion Within the Boundaries of Mere Reason* does not cover nearly enough ground to capture every variety of immoral action that human beings are capable of committing. In the last section, we saw that agents within each of Kant's categories of evil all show some level of concern for the moral standing of their actions. Both frail and impure agents do *try* to act according to the moral law, but then fail to actually incorporate the proper ordering of incentives into their maxims. Depraved agents, on the other hand, *willingly* deprioritize moral incentives in the adoption of their maxims, but only do so by *deceiving* themselves about the fact that morality commands them to do otherwise. We are therefore forced to admit that even Kant's depraved agent displays a general concern that she appears to *herself* as *morally good*.

Therefore, none of Kant's degrees of evil can explain the existence of an agent who simultaneously does all of the following:

> (1) freely chooses to prioritize non-moral incentives over the moral incentive in her maxims;
>
> (2) does not deceive herself about her own incentives for action nor about what morality commands of her;
>
> (3) does all of this without deceiving herself about the fact that her actions are immoral.

The kind of evil character exemplified by such an agent is what I henceforth refer to as *villainy*.

Intuitively, villainy seems morally worse than frailty, impurity, and depravity. Unlike frail and impure agents, villainous agents do not desire to act morally and then fail to do so in practice due to a supposed weakness of will or on account of self-deception about their true incentives for action.

Rather, villainous agents, much like depraved agents, intentionally choose to act from non-moral incentives as a matter of principle. Unlike depraved agents however, villainous agents are unfazed by the immorality of their actions, and consequently do not waste time morally justifying their actions to themselves.

Silber once argued that Hitler was something akin to a villainous agent, and that it was a defect of Kant's practical framework as a whole that it cannot give an account of his behavior (Silber 2012: 332). In what follows, I will argue that we not only have good reason to believe that villainous agents exist among us, but also, contra Silber, that they *can* be construed in a way that Kant could have accepted.

### IV.i Motivating the Need for Villainy

Before expanding upon my account of villainy and responding to several objections, it will be helpful to further motivate my claim that frailty, impurity, and depravity cannot possibly exhaust every kind of immoral action that human beings can commit. I will do this by slightly modifying a well-known example of depravity such that it reflects a degree of evil that cannot be accounted for within Kant's original three categories. In "Reflections on the Banality of (Radical) Evil," Allison gives the following example:

> "Suppose that I have a violent dislike for someone and have come into possession of a piece of information about him, which I know will cause him great pain if he learns of it. With the intent of doing so, I decide to inform him of the matter, but I justify the action on the grounds of his right to know. […] I represent it to myself (and perhaps others) as a laudable act of truth telling." (Allison 1996: 181).

Although we're already likely to view Allison's depraved agent as morally repugnant, the example can be made far worse whilst retaining its plausibility as a possible human action.

Imagine an agent in the same scenario described in Allison's example, except in this case not only does the agent decide to share the harmful information with the individual who is the object of their dislike, but also covertly spreads this information around to anyone who will listen. They do so with the malicious intention that the individual's suffering will be amplified by means of social humiliation. Moreover, let us imagine that, unlike Allison's agent, this agent does not attempt to represent her action *to herself* as a "laudable act of truth-telling" (though she may attempt to represent her action in this way *to others*). Instead, she correctly views her act solely as an attempt to inflict suffering. Let us also stipulate that this agent is unconcerned that her actions are motivated primarily by sadistic pleasure (and hence self-love), even though she would admit on reflection that doing so contradicts the moral law. Lastly, we may suppose that her lack of concern about the immorality of her actions is rooted in a general underestimation of the moral law's value.

It should be clear that none of Kant's original categories of evil can account for the kind of immoral action that is occurring in this example. This case no longer fits neatly under depravity, not even as a limiting case, since it includes neither the depraved agent's characteristic moral casuistry, nor her self-deception about what morality commands. It is also clear that it resembles neither frailty nor impurity, since it does not involve self-deception about the agent's freedom nor about her incentives. Yet, there is also nothing about this case that makes it prima facie *implausible* as a possible human action. Though the agent is admittedly devious, her action is just the consequence of her incentives from self-love when combined with her view that the moral law is not to be taken too seriously. In fact, I think that all of the features of this example could be found in a particularly vicious gossip columnist.

Although none of Kant's own categories of evil can explain this scenario, the agent in this example constitutes a paradigm case of villainy. The agent checks all of villainy's boxes:

(1) If the agent in this example consults the moral law at all, then she freely chooses to subordinate incentives from morality to incentives from self-love in the adoption of her maxims.

(2) Additionally, she clearly understands her own true incentives for action (i.e. sadistic pleasure), as well as what the moral law commands her to do in this situation.

(3) All the while, she remains undeceived about the fact that her actions are immoral.

I think that the best way to explain the actions of villainous agents within a Kantian framework is to view them as *self-deceived* about both the *source* and *value* of the moral law. That is, I believe we should view villainous agents as having deceived themselves into thinking that the moral law is merely an externally imposed set of rules for acceptable behavior (e.g. social conventions, religious dogmas, etc.) rather than an internally binding law of reason. For in doing so we can explain their serious undervaluation of the moral law's value.

To give a concrete example of this, consider someone who holds a view resembling the one put forward by Thrasymachus in Book I of Plato's *Republic*. We can easily imagine a cynical agent who believes that the moral law is simply a set of rules dictated by those who possess social power in order to serve "the advantage of the stronger" (Plato: 338 c-e). We can also understand why such an agent might be inclined to undervalue the moral law. On this agent's view, the origin of the moral law would be found in something arbitrary that exists outside of themselves, and its purpose would be to serve some other party's interests. Again, I think the best explanation for a Kantian to give when faced with an agent of this sort is to say that they have *deceived* themselves about the moral law's source, which allows them to deceive themselves about its value.

It is important to note however, that all of this is not to say that the villainous agent's self-deception about morality entails a *complete* devaluation of the moral law. Villainous agents, qua moral beings, are necessarily *aware* of the moral law's authority to some limited extent. They simply conceal this awareness under a veneer of self-deception about the moral law's source (and hence its value).

Put differently, though villainous agents *know* at some level that the moral law is authoritative for them, they deceive themselves precisely about this fact.

While I admit that it is certainly counterintuitive to claim that someone can simultaneously *know* a fact and be *self-deceived* about it, I don't believe that it is conceptually incoherent to do so. As Laura Papish points out,

> "what is decisive regarding self-deception in a Kantian framework is not whether an agent has *cognition* of what is true but whether she vigorously *attends* to the truth" (Papish 2018: 99; my emphasis).

In the case of the villainous agent, there is both an awareness of the moral law's authority and a failure to devote sufficient attention to it. This lack of attentiveness manifests in the villainous agent's ability to lie to themselves about both the source and value of the moral law, and to do so in such a way that they needn't even deny the fact that their actions contravene this law.[16] This latter ability should not surprise us, for while villainous agents are deceived about the moral law's nature, they needn't be deceived about its *content*.

Although the villainous agent's self-deception about the moral law is itself no less plausible as a piece of moral psychology than the self-deception found in Kant's original categories, there may still be some underlying concerns about the compatibility of villainy with other aspects of Kant's ethics. For example, Kant seems to have thought that agents always act under what is called "the guise of the good," i.e. that agents always represent their actions to themselves as good in *some*

[16] In the following section (IV.ii) we will see exactly how this self-deception about the moral law's source and value fits into the motivational structure of villainous action.

respect.[17] The villainous agent, by knowingly acting contrary to the moral law, may initially appear to be in violation of this principle.

Perhaps an even more concerning objection is that villainous agents seem like obvious examples of what Kant calls "diabolical beings." Again, these are beings who have completely "extirpated the dignity of the law itself" (*RGV* 6:35). Since Kant is adamant that this kind of complete perversion of morality is impossible for moral beings like us, we must ensure that villainous agents are adequately distinguished from diabolical beings. It will be the purpose of the following subsections to show first that villainy need not conflict with Kant's "guise of the good" principle, and then to show that villainous agents are not just diabolical beings in disguise. By responding to these concerns, we will also come to better understand the nature of villainy.

### IV.ii Villainy Under the Guise of the Good

Villainous agents are a profoundly irrational bunch. As we have seen, villainous agents do recognize at some level that the moral law possesses authority, since all moral beings possess the predisposition to personality, which just *consists* in recognition and respect for the moral law. In spite of this recognition however, villainous agents still choose to disregard morality's commands, and thus the dictates of practical reason, by intentionally doing what they know to be wrong. Therefore, it is tempting to argue that villainy must conflict with Kant's commitment to "the guise of the good" principle. Andrews Reath describes this principle in the following way:

> A maxim is a representation of an action or action kind as rationally supported by facts about an agent's ends, principles, circumstances, and so on, and thus represents the action *as good*. (Since the reasoning is general and

[17] Kant explicitly expresses some degree of commitment to the guise of the good thesis when he writes, "There is an old formula of the schools: [*We desire nothing except under the form of the good; nothing is avoided except under the form of the bad;*] and it has a use which is often correct," (*KpV* 5:59).

applies to all relevantly similar cases, the maxim itself is a principle to the effect that given a certain end or value and a certain set of circumstances, one is to φ, or φ-ing is good.) (Reath 2014: 11).[18]

If Reath is correct that Kant's view of maxims requires that human beings always represent their actions to themselves as good, then it might seem that the degree of clear-eyed immorality present in villainy should be impossible for Kant to accept. I think that this tension can be resolved if we explore more deeply the relationship between villainy and *moral feeling*.

Kant states that when an agent allows incentives from self-love to take unconditional precedent over incentives from morality in the adoption of their maxims, self-love is thereby transformed into *self-conceit* (KprV 5:74).[19] Both depravity and villainy can thus be thought of as constituting forms of self-conceit since they both make obedience to the moral law conditional upon its compatibility with incentives from self-love.[20] However, being consumed by self-conceit is not without consequence. When Kant says that the predisposition to personality entails *respect* for the moral law, what he has in mind is *moral feeling*. For Kant, moral feeling has two aspects: one negative and one positive. In its negative aspect, moral feeling consists in an experience of the moral law as a "humiliation of self-

[18] Reath's formulation of the guise of the good principle could be accurately described as "weak". It is weak in that it doesn't entail that representing our actions as good involves any commitment to objective moral values. Some scholars, like Korsgaard, argue that Kant endorses a much stronger version of the guise of the good thesis on which representing our actions as good requires that we take our actions to be objectively, or *morally*, good (Korsgaard 1996: 116). I think it is a mistake to attribute to Kant this "strong" version of the guise of the good principle. For a defense of this position, see Thomas Hill's book *Human Welfare and Moral Worth: Kantian Perspectives* (Hill 2002: 262-267).

[19] The exact nature of self-conceit is a matter of much controversy. One popular account explains self-conceit as the propensity to esteem oneself to such an excessive degree that you deny esteem to others. This propensity can then develop to the point where one even begins to see one's own inclinations as constituting reasons for *others* to act in ways that serve your interests (Reath 2006: 15). As an alternative to this social account of self-conceit, Owen Ware proposes that self-conceit consists in a false belief that one's own happiness is an overriding value, and in treating happiness as a "lawgiving principle." (Ware 2020: 1-4). Ware's account, I believe, better coheres with my view (defended below) that the villainous agent's self-conceit involves her taking something *other* than the moral law as a supreme source of value.

[20] It may also be the case that frailty and impurity can be thought of as forms of self-conceit. Both the frail agent and the impure agent have in fact taken up a commitment to occasionally subordinating non-moral incentives to moral ones at the level of their *supreme maxims*, even if they have adopted some lower-order maxims in which moral incentives take priority. However, it is not obvious whether Kant would take this kind of subordination of morality to self-love to be the *unconditional* subordination characteristic of self-conceit.

conceit" because it forces agents to recognize their own disordered propensity to prioritize self-love over morality (*KprV* 5:74). In its positive aspect on the other hand, moral feeling consists in an awareness of the moral law as the source of this humiliation and a feeling of respect for morality that accompanies this humiliation (*KprV* 5:79).

How does this bear on villainy? Well, despite the villainous agent's distinctive self-deception about the moral law, she still necessarily possesses the predisposition to personality. It follows then that she must experience both moral humiliation and respect for the moral law whenever she acts from self-conceit. I therefore think that one thing we must say about villainous agents is that they react perversely to these moral feelings.

Kant writes,

> "*So little* is respect a feeling of *pleasure* that we give way to it only reluctantly […] we want to be free from the intimidating respect that shows us our own unworthiness with such severity" (KprV 5:77).

Since villainous agents act from self-conceit, they too desire to be free from the painful humiliation that accompanies respect for the moral law. However, unlike most agents who set aside their self-conceit upon experiencing respect for the moral law, villainous agents should be thought of *as continuing to act from self-conceit* even in the face of moral humiliation. One explanation for why villainous agents might do this is that, in their self-conceit, they have taken the preservation of their own *pride* as the incentive to which all others are to be subordinated.[21] Put differently, they have put pride in place of the moral law as the supreme source of value.

[21] I emphasize pride for the simple reason that it provides the clearest example of how the feeling of respect might interact with self-conceit in the moral psychology of a villainous agent. However, any incentive from self-love that has been prioritized above the moral law as a source of absolute value (e.g. power or pleasure) could serve the same purpose. For, in either case, the moral feeling of respect would inflict the same humiliation upon the agent's moral self-esteem.

Through the feeling of respect, "the law […] strikes down […] pride," and therefore humbles the agent against their will (KprV 5:77). Yet, rather than doing what a more rational agent might do in this situation, i.e. subordinate their incentives from pride to moral incentives in the future so as to avoid further humiliation, villainous agents instead deceive themselves into viewing the moral feeling of respect merely as an *external infringement* upon their most cherished value (in this case, pride).

Of course, since moral feeling is a *feeling,* it must be viewed by villainous agents as an external infringement that has somehow become *internalized,* at least to the extent that it can appear within themselves *qua* feeling. To explain to themselves this internalization of something that they take to be externally sourced, villainous agents may be prone to believing that moral feeling is a mere psychological remnant from a now abandoned system of moral beliefs.[22]

What matters though for our purposes, is that in their (mis)construal of moral feeling, villainous agents must also deceive themselves about the moral law's *source* in addition to their self-deception about its *value*. For if they weren't deceived in this way, then moral feeling could never be seen by them as merely an *external* infringement on some other non-moral value.[23]

We should therefore view villainous agents as *driven* to self-deception by the promise of psychological relief from the effects of moral feeling. They believe that by deceiving themselves about the moral law they can temporarily ignore the feeling of respect (moral feeling in its positive aspect) and thereby reinterpret their moral humiliation (moral feeling in its negative aspect) as an

[22] Interestingly, Nietzsche puts forward precisely this interpretation of moral feelings in aphorism 99 of *Daybreak*: "We still draw the conclusions of judgments we consider false, of teachings in which we no longer believe — our *feelings* make us do it (Nietzsche 1997: 59; my emphasis).

[23] Again, consider an agent like Thrasymachus who tells himself that the moral law is merely an externally imposed set of rules created to serve the interests of the stronger. He would certainly not think that the moral law has its source in reason, and would therefore have nothing to prevent him from interpreting any moral feelings he might have as external infringements upon his other values.

ungrounded external imposition.[24] Then, in virtue of their self-deception about the source of the moral law, villainous agents can thereby also deceive themselves about the moral law's *value*. That is, by telling themselves the lie that the moral law is externally rather internally sourced, they can also tell themselves that the moral law is *less valuable* than whatever non-moral value they have committed themselves to, and thereby *continue* to act from self-conceit.

To sum up then, I think we should view the motivational structure of villainous action as follows:

(1.) In virtue of their self-conceit, villainous agents adopt something other than the moral law (e.g. pride) as a source of supreme value.

(2.) Accordingly, they subordinate the moral incentive to incentives from this highest value in their maxims.

(3.) On account of (2.), villainous agents experience moral humiliation and become aware of the moral law as the cause of this humiliation.

(4.) In the face of (3.), villainous agents deceive themselves into viewing the moral law as something outside of themselves in an attempt to reinterpret the moral humiliation it inspires as a mere external infringement on their preferred source of value.

(5.) The self-deception of (4.) makes possible a further self-deception about the moral law's true value, which then allows villainous agents to persist in their self-conceit.

Now let's return to where we began. If we view villainous agents as taking something like pride to be a more valuable source of incentives than morality, and as taking moral feeling to be an ungrounded infringement on that non-moral value, then we are permitted to say that they act under "the guise of the good." Recall that according to Reath, an action is regarded as "good" so long as it

[24] This is not to say that the villainous agent can ever eradicate respect for the moral law altogether, since as Kant claims, "No human being is *entirely* without moral feeling" (*MdS* 6:400; my italics). The villainous agent's erroneous view about the moral law's value will therefore *always* be inconsistent with their moral feelings.

can be, "rationally supported by facts about an agent's ends, principles, circumstances, and so on" (Reath 2014: 11). From the villainous agent's point of view, her clearly immoral actions are rationally supported by the fact that such actions serve her pride, and by the fact that moral feeling (though present) no longer acts as a *reason* to abstain from immoral action. On the contrary, her desire to avoid moral feeling has instead become a reason for her to remain in a state of self-deception about the moral law.

Although *objectively* speaking, the villainous agent's perspective involves numerous rational errors, it is still one that enables villainous agents to provide a consistent rationale for their actions *subjectively*. This kind of subjective rationale is all that Kant requires for an agent to view their own actions as being "good."

Before moving on, we should take care to point out the distinction between the sense in which villainous agents represent their actions to themselves as "good" and the sense in which frail, impure, and depraved agents represent their actions to themselves as good. As we saw earlier, agents categorizable under Kant's original three categories of evil all showed some degree of concern that their actions cohered with the moral law. Let's call this a concern that their actions are *morally good*.

Now, the primary reason that it was necessary to introduce villainy as an additional category of evil in the first place was the idea that there are plausible cases in which agents act immorally *without* any concern that their actions be *morally good*. So, it had better not be the case that what has been said above entails any concern of this kind. I see no reason why we would be forced to admit this. In fact, it would be a contradiction, given everything that's been said, to attribute any concern about *moral goodness* to villainous agents. Again, when it is said that the villainous agent sees their actions as good, all that is meant is that they see their actions as *rationally supportable* from their own *subjective* standpoint. Admittedly, this is a very weak sense of good, but we should expect no more from an agent that by definition knowingly and intentionally acts contrary to the moral law.

### IV.iii Villainy and Diabolical Beings

The second worry that one might have about my conception of villainy is that villainous agents initially appear to constitute an example of what Kant calls a "diabolical being." Recall that a diabolical being is just a being who takes defiance of the moral law as an incentive for action, and who therefore does not recognize the authority of the moral law at all. Kant has two main reasons for thinking that no human beings, with the possible exception of "morally dead" ones like sociopaths, can be diabolical beings. First, no being who does not recognize the moral law at all can be a *freely* acting being, since it is the categorical imperative that constitutes the self-legislative law necessary for freedom in the first place (RGV 6:35).[25] According to Kant, human beings are free, and therefore they necessarily acknowledge the moral law. Second, as Allison and others note, diabolical beings do not count as morally responsible *persons* since they lack the predisposition to personality altogether (Allison 1996: 177, Louden 2010: 106). Fortunately, neither of these features of diabolical beings applies to villainous agents.

First, villainous agents, unlike "diabolical beings," *do* in fact recognize the authority of the moral law, and therefore *are* able to act freely. It is not that villainous agents wholly fail to acknowledge the moral law, or even that they strive to do what is evil for evil's own sake, but rather, that they radically deceive themselves about the moral law's value relative to self-love.

I think the easiest way to understand *how* villainous agents can simultaneously recognize the authority of the moral law while also being fundamentally self-deceived about the moral law's nature is to acknowledge that there is a *positive aspect* and a *negative aspect* to their self-deception. In its positive aspect, their self-deception consists in their telling themselves the lie that the moral law is not an internally sourced law of reason but is instead something external to themselves that needn't be a supreme source of value. In its negative aspect, their self-deception consists in a lack of sufficient

[25] See the second paragraph of §II.

*attention* to their implicit awareness of the moral law's true authority, which they necessarily possess as moral beings. As we saw earlier, this negative form of self-deception does not even preclude that villainous agents *as a matter of fact* recognize that their self-deception deviates from the truth. So, while we can admit that villainous agents inadequately *reflect* on their awareness of the moral law's authority, we cannot deny them this awareness altogether.

Second, it is also not the case that villainous agents fail to possess the predisposition to personality, and therefore do not qualify as morally responsible persons. The predisposition to personality consists solely in recognition and respect for the moral law, and as we have seen, the villainous agent is capable of both. They *do* recognize the authority of the moral law, but simply do so through the distorted lens of self-deceit. Moreover, the villainous agent *does* have the moral feeling of respect, they just contort it in such a way that permits them to persist in their self-conceit.

As Kant might put it, it is not that the villainous agent, "*has* no conscience," but rather, "that he pays no heed to its verdict." (*MdS* 6:400). We might add to this that it is not that the villainous agent merely pays no heed to the verdict of conscience, but that she actively rebels against it. Nevertheless, villainous agents, unlike sociopaths or diabolical beings, *do* feel moral humiliation whenever they do what is wrong. Therefore, they share none of the essential features of diabolical beings, and deserve to be distinguished from them.

## V. The Opaque Epistemics of Action and The Aestheticization of Evil

There are two more minor objections to my view that deserve attention. Both objections were recently raised by Robert Louden in response to criticisms of Kant that depend on the claim that there exist agents who, like villainous agents, resist categorization under Kant's original taxonomy of evil. The first objection concerns the fact that the epistemics of evil action are incredibly unclear. According to Louden, since we cannot ever be certain of an agent's true

incentives, we ought to remain skeptical as to the existence of human agents who commit acts of the kind that would outstrip Kant's three categories of evil (Louden 2010: 114-115). His second objection is that even if we set this skepticism aside, we still ought to avoid accounts of evil that include this kind of agent because in doing so we would risk "aestheticizing" evil by portraying evil action as connected with the possession of a strong will or a "potent personality" (Louden 2010: 107).

I agree with Louden that we do not have perfect insight into the exact incentives that drive people to commit evil acts, but I think that it would be a mistake to conclude from this fact that we are forced into skepticism about the existence of individuals capable of villainous action. As a matter of epistemic principle, I think it is generally a bad idea to adopt skepticism in a particular domain of knowledge whenever we cannot attain *certainty* with respect to it. Instead, we ought to consult the available evidence in order to determine whether it speaks one way or another about the point in question. Bearing this in mind, and given what we know about moral psychology, it seems highly implausible that all of the evils perpetrated throughout human history can be accounted for without admitting that at least *some* were perpetrated by agents who intentionally and knowingly acted immorally.

Consider for a moment the sheer number of genocides, serial killings, and instances inhumane torture and/or execution that have occurred throughout human history. Can we really believe in good conscience that among these incorrigible acts there is not a single instance in which the agent knew that what they were doing was wrong and yet chose to do it anyways? Aren't these acts so obviously immoral that at least some of the perpetrators must've recognized the moral wrongness of their actions?

Of course, someone like Louden might argue that things are not so simple. Perhaps a sufficient number of these atrocities were perpetrated by agents afflicted by sociopathy such that we can rule

them out (since sociopathic agents would be morally inert within a Kantian paradigm). Even if we grant this, which I don't think we should, there are far more mundane cases in which agents appear to knowingly and intentionally do what is wrong. Take for example Augustine's "fruit stealing" anecdote from the *Confessions*.

> "I stole a thing of which I had plenty of my own and of much better quality. Nor did I wish to enjoy that thing which I desired to gain by theft, but rather to enjoy the actual theft and the sin of theft.
>
> In a garden nearby to our vineyard there was a pear tree, loaded with fruit that was desirable neither in appearance nor in taste. Late one night—to which hour, according to our pestilential custom, we had kept up our street games—a group of very bad youngsters set out to shake down and rob this tree. We took great loads of fruit from it, not for our own eating, but rather to throw it to the pigs; even if we did eat a little of it, we did this to do what pleased us for the reason that it was forbidden.
>
> Behold my heart, O Lord, behold my heart upon which you had mercy in the depths of the pit. Behold, now let my heart tell you what it looked for there, that I should be evil without purpose and that there should be no cause for my evil but evil itself. Foul was the evil, and I loved it." (Augustine (1960): II.iv.9).

In this passage, Augustine recalls an act of theft that he not only recognized as immoral at the time, but that he was *motivated* to commit by the mere desire to do what was wrong. Yet surely, it would be ludicrous to likewise argue that *St. Augustine* was simply a sociopath who lacked any capacity for moral reasoning.

To be clear, I don't think Augustine's act of thievery is a paradigm example of villainy. He gives us no direct evidence in the *Confessions* that he had deceived himself about the source and value of the moral law, or that he'd interpreted the moral law as an external imposition.[26] However, I don't

[26] I am even half-tempted to say that Augustine's act would constitute a *more radical* kind of evil than villainy, since it seems to lack the kind of self-deception about morality characteristic of villainy. However, doing so would pose a problem for my view given that it would strongly suggest that adding villainy to Kant's initial account would still not be sufficient to explain every kind of evil action that human beings are capable of. Therefore, I'm inclined to think it's possible for Augustine's act to be construed as a limiting case of villainy. For example, we might say that he had deceived

think that this affects the point at issue. Historical cases like Augustine's suggest that we shouldn't restrict the *potential* cases of evil action that would exceed Kant's original categories to particularly horrific acts (like Nazi atrocities), but that we should also consider everyday occurrences of immorality (like petty theft). Once we admit this, the range of possible cases in which an agent might act in such a way that would be problematic for Kant's initial account of evil expands significantly. I think this gives us very strong reason to reject Louden's skepticism about the existence of beings like villainous agents.

Louden's second objection regarding the aestheticization of evil is somewhat easier to deal with. He is primarily concerned with cases in which admitting of the existence of the kind of evil agents we've been dealing with might serve to nurture a cult of personality around such individuals. There is some support for this worry. One is reminded of fanatics who send photographs and exchange romantic letters with incarcerated serial killers, or of those who admire dictators like Hitler or Stalin for their capacity to indiscriminately impose their will on others.

Yet, I don't think that this objection poses a serious threat to my view. The villainous agent, as I have construed her, acts contrary to the moral law not out of the possession of strength of will or a "potent personality," but rather out of self-deception about the moral law's source and value. Seen in this light, villainous agents are far less conducive to being romanticized in the way Louden has in mind. They are reduced to someone who has fostered ignorance about the nature of morality due to their own desire to avoid the compunction that accompanies their selfish acts. If anything, villainous agents are more accurately viewed as moral cowards.

himself into believing that the moral law was externally imposed on him either by religious authorities (or by God), and that this allowed him to deceive himself as to the moral law's true value. Further, we might say that Augustine had also taken something else (like pride, power, or pleasure) as a supreme source of value, which in turn could explain his ability to prioritize his desire to steal over his commitment to the moral law.

Moreover, even if it were the case that villainous agents could sometimes be "aestheticized" in a way that might lead to some kind of cult of personality, I don't think that this is a good theoretical reason to abandon my view. If, as a matter of fact, there *are* plausible cases of evil action that frailty, impurity, and depravity cannot account for, and if villainy can fill this explanatory gap, then it would be a dangerous self-deception on *our* part to deny the existence of the agents capable of acting villainously. The mere fact that a truth is unsettling or dangerous is not a good reason to deny it. On the contrary, such truths deserve our attention even more urgently, for ignoring them does not make them any less dangerous, or any less true.

## VI. Conclusion

In the course of this discussion, I hope to have demonstrated the following three points:

(a.) Kant's taxonomy of evil as it is formulated in *Religion Within the Boundaries of Mere Reason* is insufficient as an exhaustive account of human immorality.

(b.) By supplementing Kant's account with villainy as an additional category of evil we can provide an intuitive explanation for the kinds of immoral actions that Kant fails to explain.

(c.) Villainy is compatible with the rest of Kant's practical framework.

If I have adequately defended these three points, then we have reason to believe that Kant's account of evil can be rendered complete by revising it along the lines that I have proposed.

In addition to this, I have pointed to a variety of self-deception that Kant failed to acknowledge. I have tried to show that it is possible for agents to become self-deceived about both the *source* and *value* of the moral law and that being so deceived does not conflict with their possessing an immediate awareness of the moral law.

As a closing remark, it may be of interest to note that there exists another form of self-deception in logical space that shares a family resemblance to the one highlighted here. It is possible to conceive of an agent who deceives herself solely about the *source* of the moral law without deceiving herself about its *value*. For example, we can imagine an agent who views morality merely as a supremely *valuable* set of social conventions rather than as a practical law of reason. I think that admitting the possibility of this latter kind of self-deception about merely the moral law's *source* may serve Kantians in explaining why so many philosophers are willing to defend non-Kantian metaethical views. However, any sufficiently detailed discussion of this would exceed the scope of this paper.

BIBLIOGRAPHY


Allais, L. (2017). Evil and Practical Reason. In E. Watkins (Ed.), *Kant on Persons and Agency* (pp. 83-101). Cambridge: Cambridge University Press.

Allison, H. E. (1996). "Reflections on the Banality of (Radical) Evil: A Kantian Analysis." *Idealism and Freedom*. Cambridge University Press.

Allison, H. E. (1990). *Kant's Theory of Freedom*. Cambridge University Press.

Allison, H. E. (2002). "On the Very Idea of a Propensity to Evil." *The Journal of Value Inquiry*, vol. 36, pp. 337-348.

Arendt, H. (1951). *The Origins of Totalitarianism*. Schocken Books.

Augustine of Hippo (1960). *The Confessions of St. Augustine* (Image Classics). Translated by John K. Ryan. Image Books.

Beck, L. W. (1960). *A Commentary on Kant's Critique of Practical Reason*. The University of Chicago Press.

Card, C. (2002). *The Atrocity Paradigm: a Theory of Evil*. Oxford University Press.

Hill, T. E., Jr. (2002). *Human Welfare and Moral Worth*. Oxford University Press.

Kant, I. (2016). *Practical Philosophy*. Translated and Edited by Mary J. Gregor, Cambridge University Press.

Kant, I. (2005). *Religion and Rational Theology*. Translated and Edited by Allen W. Wood, and Geroge Di Giovanni, Cambridge University Press.

Korsgaard, C. M. (1996). *Creating the Kingdom of Ends*. Cambridge University Press.

Korsgaard, C. M. (2013). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Louden, R. B. (2010). "Evil Everywhere: The Ordinariness of Kantian Evil." *Kant's Anatomy of Evil*. Anderson-Gold, Sharon, and Pablo Muchnik. Cambridge University Press.

Nietzsche, F. (1997) *Daybreak: Thoughts on the Prejudices of Morality*. Cambridge University Press.

Papish, L. (2018). *Kant on Evil, Self-Deception, and Moral Reform*. Oxford University Press.

Plato. (1978). *The Collected Dialogues of Plato* (Bollingen Series LXXI). E. Hamilton & H. Cairns (Eds.). Princeton University Press.

Reath, A. (2006). *Agency and Autonomy in Kant's Moral Theory: Selected Essays*. Oxford: Oxford University Press.

Reath, A. (2014). "Did Kant Hold that Rational Volition is *Sub Ratione Boni*?." https://philosophy.ucr.edu/wp-content/uploads/2014/07/Reath-Kant-on-ratl-voln.pdf. Final Draft.

Rukgaber, M. S. (2015). "Irrationality and Self-Deception Within Kant's Grades of Evil." *Kant-Studien*, vol. 106, no. 2, 234-258.

Silber, J. R. (2012). "Kant at Auschwitz." *Kant's Ethics: the Good, Freedom, and the Will*. Walter De Gruyter.

Sticker, M. (2016). When the Reflective Watch-Dog Barks: Conscience and Self-Deception in Kant. *The Journal of Value Inquiry, 51*(1), 85-104.

Timmons, M. (1994). "Evil and Imputation in Kant's Ethics." In B. Sharon Byrd, Joachim Hruschka & Jan C. Joerdan (Eds.), *Jahrbuch für Recht Und Ethik*. Duncker Und Humblot.

Ware, O. (2020). "Self-Love and Self-Conceit*." doi:https://philpapers.org/archive/WARSAS-5.pdf. Final Draft.

Wood, A. W. (1991). "Unsociable Sociability." *Philosophical Topics*, vol. 19, no. 1, pp. 325–351.