Kant's Theory of Evil: An Interpretation and Defense

by

Robert A. Gressis

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2007

Doctoral Committee:

Professor Stephen L. Darwall, Chair
Associate Professor Ian N. Proops
Assistant Professor Mika T. LaVaque-Manty
Assistant Professor Michelle A. Kosch, Cornell University

To my mother, father, and brother

Acknowledgments

I have benefited from the help of many people during the writing of this dissertation. They break down into four categories: family, friends, professors, and fiancée.

Let's talk about family first. My mother and father helped me with the dissertation by offering me unflagging love and compliments (thanks, Mom) and assurances that I'll finish the dissertation, no worries (thanks, Dad). My brother's constant headlocks, body-slams, and childish jibes reminded me that being in high school was worse than working on the dissertation. Also, a special thanks to Joey Bates (an honorary uncle) for helping me deal with "that-which-must-not-be-named" by calming me down with his wise words.

Next, friends: several graduate students in the University of Michigan philosophy department helped me get this thing done, either by making me a better philosopher (Steve Daskal, Remy Debes, Chris Dodsworth, Alex Hughes, Charles Goodman, and Steve Petersen) or by literally helping me on specific aspects of the dissertation (Remy, Chris, and Steve P.). Special kudos are owed to Steve Petersen for warning me about the mental havoc that working on a dissertation can wreak (and for helping me through it (Remy also helped with this part)) and to Chris Dodsworth for conversations helping me deal with specific problems I encountered while working through the dissertation.

My committee members all played distinct, important, and complementary roles. Steve Darwall was a chair greater-than-which-none-can-be-conceived (I prefer the cosmological argument, but here the ontological argument just works better): he kept me

thinking both of the point of the overall project and helped me immensely with particular arguments. Moreover, he was unfailingly punctual in getting comments back to me. (It's like he's done this before!)

Ian Proops was a valuable resource both for his knowledge of Kant's theoretical philosophy and for his command of style (both phraseological and sartorial!). He helped to make my writing significantly better and kept me philosophically honest.

Michelle Kosch was an encouraging and approachable interlocutor. Especially in the early stages of writing my dissertation, it was in conversation with her that I worked out most of my ideas.

Mika LaVaque-Manty generously agreed to be my cognate committee member and provided valuable suggestions about which parts of Kant's political philosophy to explore in the future.

Next, my fiancée, Shawn. Shawn has helped me in innumerable ways, in addition to being straight-up lovable. She encouraged me to keep on working when I didn't want to, she got me to use some effective time-management techniques, and she would frighten me with the specter of an unfinished dissertation, to great effect. I have matured a lot under her guidance.

Finally, I would like to give some individual laudations: thanks to David Dick, Christie Hartley, Liz Goodnick, Carole Lee, Pat Lewtas, Ira Lindsay, Pat Palermo, Matt Pugsley, Glen Shadbolt, Justin Shubow, Brent Sims, Anthony Stark, Luke Weiger, and Gabriel Zamosc-Regueros (in addition to everyone already listed) for lots of enjoyable conversations. Thanks to Chris Dodsworth for an endless parade of favors that continues to this day, and probably into the afterlife. Thanks to Project X for lots of exciting,

embarrassing adventures. Thanks to Louis Loeb and Fritz Warfield for their placement advice. Thanks to Patricia Kitcher for agreeing to share some of her precious time with me while I lived in New York City and for agreeing to write a recommendation letter. Thanks to the Knowledge Navigation Center for getting my dissertation into working order. Thanks to Dave Feldman for his counsel. Thanks to Bullett, Lennon, Muncie, and Vivi for being adorable feline beasts. And thanks again to Mom, Dad, Joe, and Shawn.

Table of Contents

List of Abbreviations

A–/B– *Critique of Pure Reason*, edited and translated by Paul Guyer and Allen W. Wood (Cambridge: Cambridge University Press, 1997).

Ant *Anthropology from a Pragmatic Point of View*, edited and translated by Robert B. Louden (Cambridge: Cambridge University Press, 2006).

CPrR *Critique of Practical Reason*, from *Practical Philosophy*, edited and translated by Mary J. Gregor (Cambridge: Cambridge University Press, 1996).

CB "Conjectural Beginning of Human History", from Toward Perpetual Peace *and Other Writings on Politics, Peace, and History*, edited by Pauline Kleingeld and translated by David L. Colclasure (New Haven, CT: Yale University Press, 2006).

CJ *Critique of the Power of Judgment*, edited by Paul Guyer and translated by Paul Guyer and Eric Matthews (Cambridge: Cambridge University Press, 2000).

E *Education*, translated by Annette Churton (Ann Arbor: The University of Michigan Press, 1960).

G *Groundwork of the Metaphysics of Morals*, from *Practical Philosophy*.

IUH "Idea for a Universal History from a Cosmopolitan Perspective", from Toward Perpetual Peace *and Other Writings on Politics, Peace, and History*.

LA *Vorlesungen über Anthropologie* [lectures on anthropology], edited by Reinhard Brandt and Werner Stark (Berlin: Walter de Gruyter & Co., 1997).

LE-H "Kant's Practical Philosophy: Herder's Lecture Notes" from *Lectures on Ethics*, edited by Peter Heath and J. B. Schneewind and translated by Peter Heath (Cambridge: Cambridge University Press, 1997).

LE-C "Moral Philosophy: Collins's Lecture Notes" from *Lectures on Ethics*.

LE-M "Morality According to Prof. Kant: Lectures on Baumgarten's Practical Philosophy", from *Lectures on Ethics*.

LE-V "Kant on the Metaphysics of Morals: Vigilantius's Lecture Notes", from *Lectures on Ethics*.

LM-D *Metaphysik Dohna*, 1792-1793 from *Lectures on metaphysics*, edited and translated by Karl Ameriks and Steve Naragon (Cambridge: Cambridge University Press, 1997). Kant's *Lectures on metaphysics* contains many Latin interpolations that I have removed for the sake of less cluttered reading. Cf. pp. xxiii.

LM-L$_1$ *Metaphysik L$_1$*, mid-1770s from *Lectures on metaphysics*.

LM-L$_2$ *Metaphysik L$_2$*, 1790-1791? from *Lectures on metaphysics*.

LM-M *Metaphysik Mrongovius*, 1782-1783 from *Lectures on metaphysics*.

LM-K$_3$ *Metaphysik Vigilantius (K$_3$)*, 1794-1795 from *Lectures on Metaphysics*.

LPDR "Lectures on the Philosophical Doctrine of Religion", from *Religion and Rational Theology*, edited and translated by Allen W. Wood and George di Giovanni (Cambridge: Cambridge University Press, 1996).

MM *The Metaphysics of Morals*, from *Practical Philosophy.*

OM "On the Miscarriage of All Philosophical Trials in Theodicy", from *Religion and Rational Theology*.

Rel *Religion within the Boundaries of Mere Reason*, from *Religion and Rational Theology*.

All italicization and boldfacing in quotations from Kant is Kant's own. All underlining is mine. I use square brackets ("[" and "]") to designate translators' interpolations and curly brackets ("{" and "}" to denote my own). Finally, when Kant writes of the *Gesinnung*, I interpolate *Gesinnung* in for the translators' "disposition", but without using curly brackets.

Abstract

Kant's theory of evil, presented most fully in his *Religion within the Boundaries of Mere Reason*, has been consistently misinterpreted since he first presented it. As a result, readers have taken it to be a mess of inconsistencies and eccentricities and so have tried to mine it for an insight or two, dismissed it altogether, or sought to explain how Kant could have gone so wrong. In this work, I provide an interpretation of Kant's theory of evil that renders it consistent and plausible.

The main problem Kant tries to solve with his theory of evil is the problem of willful immorality: how can someone who sees her moral obligations as overridingly authoritative willingly (and so culpably) transgress them? Kant's answer is that we do so by indulging various "moral fantasies" – ways of reconceiving the moral law or one's status in relation to it. By entertaining, e.g., the fantasy that one is exempt from the moral law's commands ("the exceptionalist fantasy"), or that one cannot live up to them ("despondency"), or that one need only to live up to society's standards rather than moral demands ("the adequacy fantasy") – one persuades oneself that the duty the moral law demands one to perform is only optional.

Moral fantasies appeal only because of the presence, in everyone, of the "propensity to evil," which makes the moral law's commands seem less authoritative, thereby strengthening one's sensuous desires. As a result, moral fantasies thereby diminish the urgency of moral obligations, and so tempt us to entertain them.

Scholars of Kant have missed this explanation of wickedness because they have restricted their attention to the *Religion*, a work whose gnomic statements on evil are difficult to interpret when isolated from the larger context of Kant's thinking. Assaying Kant's thinking on immorality from the mid-1780s to the late 1790s shows that Kant's main concern with evil was not so much its nature, but its very possibility. Evaluating the *Religion* in this context reveals it to be a fruitful resource for explaining the moral psychology of immorality, a problem we still deal with today, and that is possibly more relevant than ever.

Chapter 1: On *Religion within the Boundaries of Mere Reason* (A Speech to Its Cultured Despisers)

1. The Current Situation and a Historical Parallel

The title of this dissertation, *Kant's Theory of Evil: An Interpretation and Defense*,[1] is meant to echo Henry Allison's magisterial *Kant's Transcendental Idealism: An Interpretation and Defense*, for the situations of the two pieces are rather close. Even before Allison published his opus, Kant studies were thriving after a relatively long period of disdainful neglect. This was due to the groundbreaking work of P. F. Strawson (*The Bounds of Sense*) and, to a lesser degree, Jonathan Bennett (*Kant's Analytic* and *Kant's Dialectic*). Strawson and Bennett helped make Kant relevant again by showing that there was much gold in the dross of transcendental idealism; one could mine Kant for his plausible insights and discard the rest.

Allison took a different approach. Rather than using the critical philosophy piecemeal while ignoring the system, he examined how its parts fit together and then tried to take on the whole architectonic, transcendental idealism and all. In his view, the vast majority of Kant was, at the very least, worth taking seriously (if not outright correct). But the key was taking the right interpretation. As Allison saw things, people had been grievously misinterpreting Kant by reading his talk of noumena and phenomena as denoting a "real", noumenal world to which we could have no access and a lower grade, phenomenal world, knowledge of which was second-rate. Instead, we should understand Kant's talk of phenomena and noumena as referring to two different ways of

considering the same thing. This resultant reinterpretation ramified through the entirety of the critical system, making (in Allison's view, at least) transcendental idealism easier to swallow.

The theory of evil Kant presents in *Religion within the Boundaries of Mere Reason* is in a similar state to his transcendental idealism before 1983. That is, after a long period of disdainful, sometimes even vituperative,[2] dismissal, people have started taking Kant's theory of evil seriously.[3] It is not entirely clear who plays the role of Strawson (or even whether there is a Strawson) with regard to Kant's theory of evil. Arguably, that honor goes to Gordon E. Michalson, Jr., who devotes his *Fallen Freedom* to the *Religion*; like Strawson on the first *Critique*, though, he thinks much of Kant's efforts to be confused. Indeed, he thinks the *Religion* so muddled that he takes the main exegetical question to be figuring out why Kant gets himself so tangled up, rather than whether he is right.[4]

If Michalson is the Strawson of the *Religion*, then Allen Wood is undoubtedly its Allison.[5] In his *Kant's Ethical Thought*, Wood employs Kant's theory of evil as a plausible explanation of how people actually justify the ill-doing they perpetrate. At least partially as a result of Wood's work, studies of Kant's theory of evil have started to flourish (see Anderson-Gold 2001,[6] Grimm 2002, Frierson 2003, Morgan 2005, Rossi 2005, Sussman 2005, Caswell 2006a, and Caswell 2006b).[7]

In my opinion, though, both *Kant's Ethical Thought* and *Kant's Transcendental Idealism* suffer from a similar defect: they make Kant more palatable, but only at the expense of accuracy. I shall not spend any time here explaining why I do not accept Allison's "two-aspects" interpretation of transcendental idealism, but I shall explain what

is wrong with Wood's approach to the *Religion*. First, though, I want to convey why Kant's theory of evil must have looked so unappealing to so many for so long.

2. A Bad First Impression

Here are some central conclusions that Kant seems to advance in the *Religion*:

(1) Everyone is either good or evil. Consequently, no one is neutral between goodness and evil, and no one is a mixture of goodness and evil.[8]

(2) To count as good, you must be completely morally pure, i.e., you must never do anything immoral. To count as evil, you need do only one immoral thing.[9]

Given (1) and (2), it follows that:

(3) Probably everyone, at all times and places, is evil.[10]

Although (1) and (2) justify only (3), (3) is not Kant's conclusion. Instead, he holds:

(4) Everyone, at all times and places, is evil.[11]

Proposition (1) seems completely implausible. Indeed, Kant himself concedes its implausibility, writing that "Experience even seems to confirm this middle position {i.e., that people can be a mixture of good and evil or neither good nor evil} between the two extremes" (*Rel*, 6:22).

(2) is no better than (1). It would not only be absurdly demanding, but also totally removed from common usage, to describe someone who devoted her life to feeding the starving as evil if she, one time, steals some candy from a convenience store.

As for (4), it must first be said that it is a spectacularly pessimistic assessment of the species. Besides that, it simply could not be true, because we reserve "evil" to describe the worst of us, and it cannot be that all of us are the worst of us.

If Kant were to argue from (1) and (2) to (4), his argument would be plainly invalid. Fortunately, he does not do that. Unfortunately, the argument he does give for (4) is completely inadequate.

Kant accepts (4a), "everyone, at all times and places, has a 'propensity to evil'",[12] which is "the propensity of the power of choice to maxims that subordinate the incentives of the moral law to others (not moral ones)" (*Rel*, 6:30). He further seems to hold (4b), "if a person has a propensity to evil, then she is evil"; in his words, a person is evil "only because he reverses the moral order of his incentives in incorporating them into his maxims. … he makes the incentives of self-love and their inclinations the condition of compliance with the moral law" (*Rel*, 6:36). Thus, Kant's argument for (4) does not run from (1) and (2) to (4), but rather from:

(4a) Everyone, at all times and places, has a propensity to evil.

(4b) If a person has a propensity to evil, then she is evil.

to:

(4) Everyone, at all times and places, is evil.

I wrote above that Kant's argument for (4) is inadequate. Here is why. First, given the nature of the propensity to evil, we should expect an argument for (4a), the assertion that everyone has such a propensity. But Kant gives us no argument, (notoriously) claiming instead that "We can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being, in view of the multitude of woeful examples that the experience of human *deeds* parades before us" (*Rel*, 6:32-33). This will not do; the claim that everyone, at all times and places, has a propensity to evil is a synthetic, *a priori* claim; as such, it requires a transcendental argument, not mere

empirical observation. Moreover, Kant himself should have known this more than anyone, for much of the first *Critique* is predicated on the truth that we cannot prove synthetic *a priori* propositions through empirical observation.

Second, (4b) is questionable. Just because someone is disposed to commit evil does not mean that she is evil.

Kant, though, has an explanation for his take on the relationship between having a propensity to evil and being evil. The reason everyone has a propensity to evil is that:

(5) Everyone makes a free, timeless, noumenal choice to bring the propensity to evil upon herself.[13]

There are several problems with (5). First, it requires use of the phenomenal/noumenal distinction, a distinction that many find problematic. Second, it seems to assert some level of knowledge of noumena beyond that which the first *Critique* allows. Third, if everyone, at all places and times (including the future), always makes the same choice—namely, to bring upon herself a propensity to evil—then it is hard to see how this choice could be a free one. Instead, it seems to make the propensity to evil necessary. But if the propensity to evil is necessary, then it could not be the result of a free choice.

Even if this choice were neither metaphysically necessary nor one everyone made, there would be another huge problem with it. For Kant is clear that:

(6) The only reason anyone is evil is that she chooses to make herself evil.

We can elicit a problem with (6) by asking the question, What is the reason people have for making themselves evil? If one says that people have no reason, then this appears to commit Kant to the possibility of whimsical, reasonless choices. However, and as I show

in chapter 2, he thinks all action by any finite person is an action on a maxim.[14] Thus, his action-theory forbids him from saying this.

More fundamentally, one cannot say that the reason a person makes herself evil is that she is evil, because she is not evil until she makes this decision. However, if one claims that she was not evil before making this choice, then it becomes impossible to explain why she should be counted as evil after making the choice. This is because we cannot "accidentally" make evil choices – if I do something thinking it morally innocent, or because I genuinely misunderstand my moral obligation, then even if it has terrible results, it cannot be called evil. As Kant puts it:

> Evil can have originated only from moral evil (not just from the limitations of our nature); yet the original predisposition (which none other than the human being himself could have corrupted, if this corruption is to be imputed to him) is a predisposition to the good; there is no conceivable ground for us, therefore, from which moral evil could first have come in us. (*Rel*, 6:43)

The sentiment, "Evil can have originated only from moral evil" is right, but the claim that the origin of evil is inconceivable is (or at least, seems to be) wrong. It is not that the origin of evil seems to be inconceivable; rather, given Kant's views, it seems to be impossible. This is because as purely noumenal beings, we have no sensible desires; if we have no sensible desires, though, then the only ground available to us, the only principle in virtue of which things can even count as reasons for action, is the moral law.[15] But a being who guides herself exclusively by the moral law cannot possibly act against that law.

Kant does not see the inconceivability of (6) as a problem; instead, he seems to use it as an inspiration for another of his doctrines, that of moral revolution:

(7) Everyone is morally obligated to become good.[16]

When we add to (7) the fact that Kant endorses the principle, "'ought' implies 'can'",[17] then (7) ends up contradicting (4); if everyone is, at all times and places, evil, then one cannot ever be good (unless (1), Kant's rigoristic principle that people cannot be both good and evil, is false). But if it is impossible for a person to become good, then she cannot be morally obligated to try to become good.

Why then, does Kant advance as a moral demand something even he immediately recognizes cannot be satisfied? Regrettably, Kant sees the inconceivability involved in asserting (6), not as itself a problem, but rather as a justification for (7):

> How it is possible that a naturally evil human being should make himself into a good human being surpasses every concept of ours. … But … since the fall from good into evil (if we seriously consider that evil originates from freedom) is no more comprehensible than the ascent from evil back to the good, then the possibility of this last cannot be disputed. (*Rel*, 6:44-45)

In the face of all these extraordinary claims, many of which contradict each other, one could simply throw up one's hands and say, "I grant that most of the *Religion* is confused. But what really matters is not the metaphysical or physiological grounding of evil; what matters to us nowadays is how Kant understands evil action – both what it is and what motivates it."

Unfortunately, even a person scouring the *Religion* just for Kant's understanding of evil action will be disappointed, because Kant's understanding of the nature of evil seems to be so simpleminded. It appears to be Kant's view that:

(8) The only reason anyone acts evilly is to make herself happier.

The problem with (8) is that it makes evil too simple and consequently, too common. First, it makes evil too simple: there is a great variety of evil action, but reducing it all to the quest for personal pleasure makes it difficult to understand actions that are clearly evil but are also clearly not done just for pleasure, such as suicide bombing. Second, it makes

evil too common: because any subordination of morality to happiness counts as evil, we have the result that something as minor as stealing a candy bar ends up being evil. We do not describe any such act as "evil", though; rather, we judge only the worst instances of immorality (such as wanton murder) to be evil. Thus, Kant uses promiscuously a concept that most of us reserve for special occasions.

Confronted with this picture, it is no wonder that most Kantians would rather ignore it.[18] It appears to be a welter of confusions, contradictions, and eccentricities.

3. Wood's Advance

In light of these apparently immense difficulties, Wood steps in with an ingenious interpretative move: having a propensity to evil is identical to being "unsociably sociable".[19]

On the theory of unsociable sociability, what causes society to advance and develop – competition among people for social status – is also what causes people to dislike each other and act immorally.[20] The idea of unsociable sociability is this: each of us naturally wants social approval and status (respect from our peers, wealth, power); to get this, we will do all sorts of things to improve ourselves – this is the "sociability" part. At the same time, though, the root of our drive for self-improvement is a false standard of worth. According to the "rational-moral"[21] standard of the moral law, everyone has equal dignity, and gains moral credit the more closely she lives up to its standards; according to what Wood calls the "natural-social"[22] standard, though, the better you are at something relative to others, the better a person you are. Since people are deeply concerned about their happiness,[23] and since the largest portion of their happiness comes from how they

think they compare to others,[24] people are inclined to subordinate morality to the promotion of their comparative worth.

On Wood's view, then, what gives people a propensity to evil is not the making of a free, timeless, noumenal choice, but rather growing up in a society that relies on the natural-social standard of evaluation as its main standard of moral evaluation. This allows Wood to avoid the metaphysics of (5) ("everyone makes a free, timeless, noumenal choice to bring the propensity to evil upon herself") and the absurdity of (6) ("the only reason anyone is evil is that she chooses to make herself evil").

Moreover, equating the propensity to evil with unsociable sociability also allows Wood to give a justification for (4a) ("everyone, at all times and places, has a propensity to evil"). Since, according to Kant, nature (or providence) "wants" the human species to develop its predispositions to culture and morality,[25] and because unsociable sociability is the mechanism nature uses to do this,[26] it follows that Kant's justification for (4a) (and, consequently, for (4) ("everyone, at all times and places, is evil")) stems from his theory of history. Thus, neither (4) nor (4a) is an assertion with a mere patina of empirical support, but is instead rooted in a well-worked out theory of historical development.

Arguably, though, Wood's greatest achievement (and the portion of Kant's theory of evil to which he devotes the most attention)[27] is to bring Kant's conception of evil closer to our own. Most people do not think of evil as having to do just with selfishness; malice, hatred, callousness, and other such attitudes are more characteristic of evil than is egoism. Wood shows how the desire for happiness – in particular, the part of happiness having to do with thinking of oneself as better than others – leads to all manner of immoral behavior. This permits Wood to refrain from attributing (8) ("the only reason

anyone acts evilly is to make herself happier") to Kant. It also allows him to go some way towards making sense of (1) ("everyone is either good or evil" with no middle ground): either you value yourself primarily according to the rational-moral conception (in which case you are good) or to the natural-social conception of value (in which case you are evil).

Wood's interpretation is a valiant effort. It does not, though, vindicate all the central claims of Kant's theory of evil. For example, although (1) becomes more plausible on Wood's view than it was before, it is still too extreme; there seems to be nothing intrinsically impossible with someone using the rational-moral standard to evaluate some things and the natural-social standard to evaluate others, without making either primary.

Even if it is the case, though, that Wood saves (1), he still does not save (4) ("everyone, at all times and places, is evil"); even if the natural-social standard of evaluation is the socially dominant standard, there could still be some dissidents who use the rational-moral one. Thus, Wood at best gets, "most people, at all times and places, are evil"; which is not Kant's actual conclusion. (Nonetheless, it would be a very significant result.)

Arguably, then, Wood does not save (1) or (4). But there are other claims he does not even try to save, such as (2) ("to count as good, you must be completely morally pure, i.e., you must never do anything immoral. To count as evil, you need do only one immoral thing") and (6) ("the only reason anyone is evil is that she chooses to make herself evil"). (2) is perhaps a claim best abandoned, but (6) seems to be rather central to Kant's account of moral responsibility; if you are evil—if, that is, you make the natural-

social standard of evaluation primary—then it cannot be the case, at least not for a Kantian, that your society is fully responsible for making you evil. Unless you share at least some of the blame for your evil, only your society will count as evil.[28]

Moreover, Wood does not address much attention to Kant's doctrine of moral revolution. Admittedly, if people are not necessarily evil – and on Wood's interpretation they are not – then our being obligated to become good is now perfectly straightforward. But this clearly departs from Kant's view of the matter, for he sees the change from evil to good to be inconceivable.

This actually gets to the heart of the problem for Wood. What is lacking in Wood's approach is not that he does not salvage every outrageous claim of Kant's, but rather that his central interpretative move – that the propensity to evil is just unsociable sociability – is wrong. I sketch the reason why in the next section, but I should note that regardless of his interpretation's faults, it still marks the most valuable contribution to the study of Kant's theory of evil we so far have.

4. A Tour of What Is to Come

I should say at the outset that though this dissertation is intended to be an interpretation and defense of Kant's theory of evil, limitations of time have prevented me from defending as much of Kant's theory as I would like. What I do instead is, mostly, clear the ground for a defense by providing what I think is the right interpretation. More precisely, I show that Kant's theory of evil is not as hard to defend as people think, for he does not make claims nearly as unbelievable as the ones people attribute to him.

To start things off, in chapter 2 ("Maxims and *Gesinnungen*") I explore Kant's theory of maxims. I do this for two reasons: first, to understand how Kant thinks evil

actions work, you have to have at least some understanding of how he thinks actions in general work. And as it turns out, crucial to understanding how actions work is understanding his notion of maxims; for on Kant's view, what separates a mere physical event from a physical event that is also an action is that the physical events that are actions are motivated by maxims. That is, every action is action on a maxim (I call this the "Maxim-Action Thesis").

The second reason I explicate Kant's theory of maxims is that central to his theory of evil is the notion of an evil *Gesinnung*, or "supreme maxim". I end up showing, first, what evil *Gesinnungen* are: namely, maxims to subordinate one of an agent's highest maxims (either the maxim of allegiance to the moral law, which I dub the "Moral Maxim", or the maxim of promoting one's own happiness, which I name the "Prudential Maxim") to the other. And I also show how evil *Gesinnungen* work: by making certain sensible desires appear to have more going for them, all things considered, than certain moral obligations.

In chapter 3 ("The Propensity to Evil and the *Gesinnung*"), I begin to get at the heart of Kant's theory of evil. This I do by exploring the relationship between the evil *Gesinnung* and the propensity to evil. On my interpretation, the propensity to evil is what Kant calls a "contingent predisposition" (*Rel*, 6:28), that is, a predisposition to develop certain kinds of desires, but that is also not part of the concept of the human person (unlike, say, the predisposition to acting from respect for the law, which is tied up with the concept of a person).[29] That is, it is a property the possession of which renders one susceptible—given the right triggering conditions—to developing a standing desire to engage in evil action. In contemporary parlance, it is a second-order disposition, i.e., a

disposition that, given the right triggering conditions, gives rise to a disposition to act immorally.

Analyzing the propensity to evil as a contingent predisposition allows us to separate the propensity into two elements. First, what I call a "susceptibility to evil", which is the propensity to evil considered only as a person's susceptibility to develop a standing inclination to evil if put in the right conditions; and second, a "tendency to evil" (my title), which is the standing inclination to which the susceptibility can give rise. When Kant talks about the propensity to evil, he almost always means the tendency to evil.

Since the propensity to evil is contingent, it follows that we can imagine a situation in which not everyone has it. This situation is a world exemplifying the highest good.[30] But if there is a possible state of affairs in which the propensity to evil (considered, specifically as a tendency to evil) is not possessed by anyone, then people are wrong to see (4a) ("everyone, at all times and places, has a propensity to evil") as a claim Kant makes. Thus, we do not need to find a transcendental argument for (4a), because Kant does not make the synthetic, *a priori* claim that <u>everyone</u> has a propensity to evil.[31] Rather, he claims:

(4c)Everyone in all evil societies[32] has a propensity to evil (*qua* tendency to evil).

So Kant does not accept (4a). Just as important, though, he does not believe (4b) ("if a person has a propensity to evil, then she is evil") either. He does not believe that merely having a propensity to evil makes a person evil; instead, he holds:

(4d)<u>Having an evil *Gesinnung*</u> makes a person evil.

That is, a person is evil not because she has a standing desire to satisfy her sensible desires rather than her moral obligations; rather, she is evil because she actually subordinates her moral obligations to her sensible desires (by subordinating the Moral Maxim to the Prudential Maxim).

Consequently, Kant does not believe (4) ("everyone, at all times and places, is evil"); instead he believes:

(4') Everyone who has an evil *Gesinnung* is evil.

Looking just at (4'), one has no idea how widespread human evil is. As it turns out, Kant does believe evil to be ubiquitous; this is because he thinks that, in order to activate a propensity to evil *qua* tendency to evil, one needs to adopt an evil *Gesinnung*. Since everyone in every evil society has an active propensity to evil, though, does it not follow that Kant thinks, after all, that all of us (except those happy few in the highest good) are evil, just as the standard interpretation has it?

No. This is because Kant believes that, while the evil *Gesinnung* is necessary to activate the propensity to evil, once activated that propensity stays activated, regardless of whether one's evil *Gesinnung* remains. So an evil *Gesinnung* is needed to start the ball rolling with regard to the propensity to evil, but once it does that, the propensity to evil continues rolling along on its own. Thus, though everyone in an evil society has a propensity to evil, not everyone in an evil society has an evil *Gesinnung* – i.e., not everyone in evil societies is evil.

If not everyone in evil societies is evil, about how many are? In Kant's view, most of us are evil. Some of us, though, succeed in having a moral revolution and becoming good (i.e., taking on a good *Gesinnung*).[33] So, Kant really believes (7) ("everyone is

morally obligated to become good"); only now, because (7) no longer conflicts with (4) ("everyone, at all times and places, is evil"), it is possible to accept.

In chapter 3 I show that Kant does not believe (4a) ("everyone, at all times and places, has a propensity to evil") but rather believes (4c) ("everyone in all evil societies has a propensity to evil (*qua* tendency to evil)"). While (4c), a claim I call the "Universality Thesis", is more plausible than (4a), it still requires an argument. I provide this argument in chapter 4 ("Kant's Argument for the Universality Thesis").

Since Kant claims only the people in evil societies have a tendency to evil, he cannot give a transcendental argument for the claim. What sort of argument, then, can he give for it? I argue that we cannot easily find the argument by looking at the *Religion*; instead, we have to investigate Kant's earlier (1783-86) "Lectures on the Philosophical Doctrine of Religion" and the Collins manuscript of his 1784-85 lectures on moral philosophy (which became the *Groundwork of the Metaphysics of Morals*).

In both his lectures on religion and his lectures on ethics, Kant denies the existence of a propensity to evil. In the theological lectures, Kant does not explicitly give a reason for denying a propensity to evil, but his worry seems to be that if we did have a propensity to evil, God would be morally responsible for our having it. Consequently, any evil anyone commits results either from ignorance or weakness of will.

In the lectures on ethics, Kant denies the propensity to evil, not because it would make God culpable for evil, but because it conflicts with his theory of historical development. In particular, having a propensity to evil would conflict with Kant's theory of unsociable sociability; this is because the theory of unsociable sociability is supposed to explain how human societies advance and develop the human predispositions to

culture and morality. If the propensity to evil existed, though, people would engage in wholeheartedly evil acts, which Kant calls "devilish vice".[34] The problem with allowing for wholeheartedly evil acts, though, is that they undermine the development of the human predispositions – consequently, it would thwart nature's "purposes" to "allow for" a propensity to evil.

It is important to note that in both Kant's lectures on religion and on ethics, Kant denies a propensity to evil for more or less the same reason. In the theological lectures, Kant must rule out a propensity to evil because such a propensity would allow for evil acts born neither of ignorance nor weakness of will, but rather stemming from a "*special germ toward evil*" (*LPDR*, 27:1078); if God allowed such acts, we would have to call his goodness into question. Similarly, in the lectures on ethics, Kant cannot allow for a "direction inclination to evil" (*LE-C*, 27:441) (i.e., a propensity to evil) because such a propensity would allow for wholeheartedly evil acts, which would undercut unsociable sociability – the engine providence uses to bring about its goals.[35] Thus, if providence allowed for a propensity to evil, we would have cause to doubt its goodness.

Given the religious rationale behind Kant's denials of a propensity to evil, it makes sense to look at his later (1791) essay, "On the Miscarriage of All Philosophical Trials in Theodicy", wherein he rejects his old theodicy and advances a new one,[36] to see whether he also rejects his old views on the propensity to evil. And what one finds is that Kant does indeed assert a propensity to evil (which he calls a "propensity to mendacity") there. In other words, because he rejects his old theodicy, Kant also must give up his old denial of a propensity to evil. Consequently, with "On the Miscarriage …" Kant allows for the possibility of wholeheartedly evil actions.

Wholeheartedly evil actions are now possible; but how does this relate to Kant's "formal proof" of the Universality Thesis? In this way: Kant's explanation for how wholeheartedly evil actions are possible is the propensity to evil *qua* tendency to evil. In other words, having a tendency to evil is what makes one capable of engaging in wholeheartedly evil action. So, if you are sure that everyone in evil societies is capable of engaging in wholeheartedly evil action, you can be sure that everyone in such societies has a propensity to evil. Kant's argument for the Universality Thesis, then, is:

(9) Everyone in an evil society is capable of wholeheartedly evil action.

(10) If a person is capable of wholeheartedly evil action, then she has a propensity to evil.

(11) Therefore, everyone in evil societies has a propensity to evil.

The main problem with this argument is: how can you be sure that everyone is capable of wholehearted evil? As has been noted, Kant does not offer an argument for this claim, but instead makes a few empirical observations. The reason he makes empirical observations, though, is that he thinks that a denial of (10) is itself evidence that one has a propensity to evil; that is, if you are eager to say that all evil actions stem either from ignorance or weakness of will, it is probably because you want to exonerate yourself (or someone you care about). If this is true, then there would be no point in offering an argument to such a person, because she shows herself to be arguing in bad faith. The best you can do with someone so recalcitrant is simply show her acts of evil, and then ask whether she <u>really thinks</u> such an act stems from ignorance or *akrasia*.

Kant does not just sit back and take for granted that wholeheartedly evil actions are possible, though. He offers an elaborate theory of how people can end up performing

such actions, a theory I explicate in chapter 5 ("How to Be Evil: Kant's Moral Psychology of Immorality").

Kant needs to explain the possibility of wicked actions, because there are powerful reasons, stemming from some central doctrines of his ethical theory, which make their possibility doubtful. Let me clarify: everyone knows that the reasons offered by the moral law must override the reasons for action provided by her desires;[37] moreover, a person cannot be ignorant of this fact, because her conscience will make the moral law's deliverances vivid to her, especially when she considers acting against the moral law.[38] Given the intrusions of the moral law through conscience, and everyone's assent to the overridingness of the moral law's demands, how can people nonetheless ignore those demands and engage in willful evil?

Kant's answer is: we tell ourselves stories. According to these stories, something about us makes us special, and it is because we are special in this way that we are permitted to violate the moral law that would normally apply to us. In other words, we tell ourselves stories about ourselves, and when we believe the stories, we end up with evil self-conceptions according to which it is morally permissible, at least for us, to do wrong.

This answer might seem unsatisfactory. After all, how do we manage to get ourselves to believe these stories, given the moral law's constant loitering in our minds? According to Kant, we do so by employing particular techniques to befuddle our consciences: when conscience shows up, we either rationalize our behavior to ourselves, or we change the subject to stop its torment.

This just pushes the question one step further back, though. For if we are not morally responsible for employing these techniques of conscience-bamboozlement, then we cannot be held responsible for whatever evil results from them; but if we are morally responsible for them, then we have to ask: how do we manage to convince ourselves that it is okay to use the techniques?

Here Kant has to dig in his heels: we are capable of convincing ourselves about the techniques' propriety because we start out with an evil *Gesinnung* according to which those techniques are morally permissible. It is because of this initial evil *Gesinnung* that we can make ourselves into people in whom the evil *Gesinnung* fully blossoms. In other words, Kant believes (6) ("the only reason anyone is evil is that she chooses to make herself evil"), and the incomprehensibility, perhaps even impossibility, this entails.

I do not think, though, that this problem—that the willfully evil action is possible only if one is already assumed to be evil—shows willful evil to be impossible. Instead, I agree with Kant that the possibility of evil action is, at worst, inconceivable. I agree with him for two reasons: first, the problem of willful evil arises for anyone who believes both that moral reasons are overriding and that truly evil action requires a direct confrontation with and subordination of moral reasons to reasons even the agent knows, on some level, to be less weighty. Since I think many of us have strong intuitions both about the sovereignty of moral reasons and about the necessary conditions of evil action, it turns out that many of us are in the same boat Kant is. Indeed, I am not even sure that I could give up either intuition; and if I cannot, what should I do other than assert them both, admit the inconceivability of their joint truth, and hope that nonetheless they are both correct?[39]

The second reason I find evil action to be only inconceivable, rather than impossible, is that I disagree with the standard take that we, as noumenal beings, cannot act evilly. Moreover, I do not think that Kant believes this either. While it is true that beings who were only noumenal could act only according to the moral law, we are not such beings. We are both noumenal and phenomenal; this means that we noumenally bring about effects in the phenomenal world. But because we are also phenomenal beings, we can bring sensible influences to bear on how we make those noumenal decisions. Thus, there is nothing intrinsically impossible in the notion of a free, timeless, evil noumenal decision.[40]

Thus, while I admit that (6) ("the only reason anyone is evil is that she chooses to make herself evil") is inconceivable, I do not think it impossible. Moreover, while I find (5) ("everyone makes a free, timeless, noumenal choice to bring the propensity to evil upon herself") to be somewhat metaphysically extravagant, I think the fact that we have to endorse (6) makes (5) as good an explanation of (6)'s truth as any other. That is, if you admit a choice to be inconceivable, why not root its happening in a realm about which it is impossible for you to know much?[41]

So in chapter 5 I show both (5) and (6) to have more going for them than other interpreters do. But I do not take my defensive maneuvers with regard to (5) and (6) to be the main accomplishment of chapter 5. I take that to be my display of the power of Kant's theory of self-conceptions for explaining evil actions, despite Kant's embrace of hedonism about non-moral value and motivation. That is, even though I do not, ultimately, accept (8) ("the only reason anyone acts evilly is to make herself happier"), I think people radically oversimplify it; once one realizes the degree of explanatory power

Kant has even in spite of (8), one has reason to find his theory of evil plausible and illuminating, after only minor tweaking.

I take Kant's view of evil self-conceptions to be this: people adopt certain evil self-conceptions because they think (on some level) that doing so will make them happier. Once a person has adopted an evil self-conception, then what she counts as making her happier changes from her preferences before she adopted that self-conception. After someone dons some evil self-conception *E*, actions and results that confirm she has *E* are what make her happier. So Kant's theory still suffers from the defect that an evil person, when she is not acting morally, acts only for the sake of pleasure; however, the things that give her pleasure will be intimately tied up with her identity, so Kant's hedonism ends up being more plausible than it first seems.

5. Our Location at the End of the Tour

In section 2 of this chapter, I listed several outrageous claims, some of which seem to be inconsistent. They were:

(1) Everyone is either good or evil. Consequently, no one is neutral between goodness and evil, and no one is a mixture of goodness and evil.

(2) To count as good, you must be completely morally pure, i.e., you must never do anything immoral. To count as evil, you need do only one immoral thing.

(4) Everyone, at all times and places, is evil.

(I also noted that Kant is standardly thought to justify (4) with (4a) ("everyone, at all times and places, has a propensity to evil") and (4b) ("if a person has a propensity to evil, then she is evil").)

(5) Everyone makes a free, timeless, noumenal choice to bring the propensity to evil upon herself.

(6) The only reason anyone is evil is that she chooses to make herself evil.

(7) Everyone is morally obligated to become good.

(8) The only reason anyone acts evilly is to make herself happier.

In section 4 of this chapter, I showed that Kant does not actually claim (4) (instead, he holds that "everyone, in all evil societies, has a propensity to evil"), or (4a) (because people in the highest good do not have a propensity to evil *qua* tendency to evil), or (4b) (instead, he holds that people with evil *Gesinnungen* are evil – but having an evil *Gesinnung* is not the same as having a propensity to evil). Moreover, I showed that while (6) brings some absurdity with it, arguably any position one takes relating to (6) brings absurdity with it; and since (6) is already inconceivable, there is nothing particularly questionable about explaining (6) via recourse to (5). In addition, I showed that (7) is perfectly acceptable, because Kant does not actually assert (4), the claim with which it earlier conflicted. As for (8), I conceded that it is probably not right, but that something very close to it could very well be right.[42]

This leaves (2). I happen to think that (2) has a lot going for it, but unfortunately I do not have space in this dissertation to argue for this. In a nutshell, though, Kant's complex notion of "purism"[43] is quite possibly right, at least if one accepts a demanding moral theory like Kant (or Peter Singer or Peter Unger)[44] does. But showing that is a task for a later time.

---

[1] Thanks to Ian Proops for the title.

[2] I have in mind here Goethe's famous remark to Herder that "Kant required a long lifetime to purify his philosophical mantle of many impurities and prejudices. And now he has wantonly tainted it with the

shameful stain of radical evil, in order that Christians might be attracted to kiss its hem" (Fackenheim 1954, 340).

[3] I would say they have started taking it seriously "again", but it not clear that people *ever* took it seriously before Wood 1999.

[4] See Michalson 1990, 1-10. See also Bernstein 2002 for a scholar who takes a similar tack.

[5] Actually, "undoubtedly" might be a bit too strong, as Allison himself has a strong claim to be considered the Allison of the *Religion*. Allison defends what Kant has to say in the *Religion* in Allison 1990, Allison 1996, Allison 2001, and Allison 2002, but he seems more interested in showing Kant to avoid inconsistency than to show him to be correct.

[6] In fairness to Anderson-Gold, her Anderson-Gold 1991 itself inspired Wood; see Wood 1999, 287.

[7] In fairness to Allison, Morgan 2005, Caswell 2006a, and Caswell 2006b seem to hew more closely to his approach than to Wood's.

[8] "*The human being is* (by nature) *either morally good or morally evil*" (*Rel*, 6:22).

[9] "In order … to call a human being evil, it must be possible to infer *a priori* from a number of consciously evil actions, or even from a single one, an underlying evil maxim, and from this, the presence in the subject of a common ground, itself a maxim, of all particular morally evil maxims" (*Rel*, 6:20).

[10] Engstrom 1988 sees (1) and (2) in Kant, and judges that they lead to (3). He calls (1) "rigorism" and (2) "purism" (Engstrom 1988, 436).

[11] We "may presuppose evil as subjectively necessary in every human being, even the best" (Rel, 6:32). See also *Rel*, 6:30 and 32-33.

[12] See *Rel*, 6:32-33.

[13] The:

> good or the evil in the human being is said to be innate (as the subjective first ground of the adoption of this or that maxims with respect to the moral law) only *in the sense* that it is posited as the ground antecedent to every use of freedom given in experience (from the earliest youth as far back as birth) and is thus represented as present in the human being at the moment of birth (*Rel*, 6:21-22).

[14] I write "finite person" because God is a person who does not act on maxims: "All three concepts, however – that of *incentive*, of an *interest* and of a *maxim* – can be applied only to finite beings" (*CPrR*, 5:79). However, God does not act whimsically but instead always on the moral law: "no imperatives hold for the *divine* will and in general for a *holy* will: the 'ought' is out of place here, because volition is of itself necessarily in accord with the law" (*G*, 4:414).

[15] "All my actions as only a member of the world of understanding would therefore conform perfectly with the principle of autonomy of the pure will" (*G*, 4:453).

[16] If "a human being is corrupt in the very ground of his maxims, how can he possibly bring about this revolution by his own forces and become a good human being on his own? Yet duty commands that he be good, and duty commands nothing but what we can do" (*Rel*, 6:47).

[17] The "subjective effect of this {moral} law, namely the disposition conformed with it and also made necessary by it to promote the practically possible highest good, nevertheless presupposes at least that the latter is *possible*; in the contrary case, it would be practically impossible to strive for the object of a concept that would be, at bottom, empty and without an object" (*CPrR*, 5:143).

[18] Seiriol Morgan writes:

> Even more telling, perhaps, than the almost uniform verdict of failure handed down by those who have explicitly addressed the claims of the Religion is the widespread silence on the matter from those providing exegeses or defenses of Kantian moral philosophy in general. Clearly, if true, his assertions regarding a universal human propensity to evil could not but have significant implications for moral psychology and ethics in general. And yet numerous standard twentieth-century textbooks on Kant's moral philosophy say nothing at all on the matter. (Morgan 2005, 63-64)

[19] See Wood 1999, 334.

[20] *IUH*, 8:20.

[21] Wood 1999, 241-42.

[22] Wood 1999, 241.

[23] "{A}ll people have already, of themselves, the strongest and deepest inclination to happiness" (*G*, 4:399).

[24] "The greatest source of happiness or unhappiness, of faring well or ill, of content or discontent, lies in the relationship to other people" (*LE-C*, 27:366-67).
[25] "*All of a creature's natural predispositions are destined eventually to develop fully and in accordance with their purpose*" (*IUH*, 8:18).
[26] "Without those characteristics of unsociability … human beings would live the arcadian life of shepherds, in full harmony, contentment, and mutual love. But all human talents would thus lie eternally dormant" (*IUH*, 8:21).
[27] See Wood 1999, 250-320 and Wood 1996.
[28] Wood does respond to this in his later *Kant*: "Acting from our propensity to unsociable sociability is something we do freely, and for which we are to blame" (Wood 2005, 118), but this response misses the point. Kant does not think that we are responsible just for the individual evil actions we perform because we have a propensity to evil; he thinks we are responsible for having a propensity to evil in the first place.
[29] Kant calls the predisposition to act from respect for the moral law the "predisposition to personality" (*Rel*, 6:27-28).
[30] See *Rel*, 6:94.
[31] Strictly speaking, Kant *does* believe that everyone has a propensity to evil considered just as a susceptibility to evil. But when people attribute (4a) to Kant, they always understand "propensity to evil" in the way I understand "tendency to evil". And Kant does not believe that everyone, in all times and places, has *that*.
[32] I take an evil society to be any society that does not exemplify the highest good. So this includes almost every society that ever was or will be.
[33] See *Ant*, 7:294, which shows only a few of us to succeed in carrying out this moral revolution.
[34] See *LE-C*, 27:380 and 439.
[35] See *IUH*, 8:30.
[36] For an excellent exposition of the details of that theodicy, which I do not examine in detail in this dissertation, see Pereboom 1996.
[37] The "necessity of my action from *pure* respect for the practical law is what constitutes duty, to which every other motive must give way because it is the condition of a will good *in itself*, the worth of which surpasses all else" (*G*, 4:403).
[38] "Every human being has a conscience and finds himself observed, threatened, and, in general, kept in awe (respect coupled with fear) by an internal judge … It follows him like his shadow when he plans to escape" (*MM*, 6:438).
[39] Peter van Inwagen finds himself in a similar position regarding the dilemma of determinism: he finds he cannot believe compatibilism, but also that he cannot be a skeptic about moral responsibility. Thus, he believes in both incompatibilism and free will; but he cannot see how they could both be true. However, he finds the "mystery" of incompatibilistic free will to be a "smaller" mystery than either compatibilism or the impossibility of moral responsibility. See van Inwagen 2002, 215-16. Arguably, I could make the same move: the idea that there is, in fact, no immorality except that done from weakness or ignorance, or the idea that moral obligations are not in fact overriding is, for me anyway, harder to swallow than the idea that there could be willful evil.
[40] I do not elaborate upon this view in the rest of the dissertation.
[41] At least, if you believe in such a realm in the first place.
[42] Indeed, one social psychological approach to explaining evil – that of Crocker, et al. 2004 – dovetails quite closely with Kant's own self-conception approach to evil.
[43] See endnote 10.
[44] See Singer 1995 and Unger 1996.

Chapter 2: Maxims and *Gesinnungen*

1. Introduction

In this chapter I explain the relationship between maxims and the *Gesinnung*. While maxims are subjective practical principles, a person's[1] *Gesinnung* is her "supreme maxim" (*Rel*, 6:31 and 66), the adoption of which influences the adoption of all her other maxims.[2] To clarify this relationship, it is necessary to define both maxims and the *Gesinnung* in greater detail. Thus, this chapter has three goals: first, to describe what a maxim is; second, to explain what a *Gesinnung* is; and third, to articulate the relationship between them.

2. Maxims

Kant uses "maxim" to refer to two different kinds of thing. First, maxims are motivating judgments of some state of affairs or action's goodness or badness/evil (throughout the rest of this chapter, I shall leave out the "badness/evil" part and take maxims to be motivating judgments just of a state of affairs or action's goodness). Second, maxims are formulations of general policies, either of thinking or acting. I call the first sort of maxim a "disposition-maxim" and the second sort a "policy-maxim".

Unfortunately, Kant never says that he uses maxim to name two different types of thing. Instead, when he defines maxims, he gives the impression that he uses the term univocally. The following definitions, which span from 1785 to 1797, are representative:

> [M1] A *maxim* is the subjective principle of acting, and must be distinguished from the *objective* principle, namely the practical law. The former contains the practical rule determined by reason conformably with the conditions of the subject (often his ignorance or also his inclinations), and is therefore the principle in accordance with which the

subject *acts;* but the law is the objective principle valid for every rational being, and the principle in accordance with which he *ought to act*, i.e. an imperative. (*G*, 4:421n)

[M2] Practical *principles* are propositions that contain a general determination of the will, having under it {i.e., the determination} several practical rules. They are subjective, or *maxims*, when the condition is regarded by the subject as holding only for his will (*CPrR*, 5:19).

[M3] A *maxim* is a *subjective* principle of action, a principle which the subject himself makes his rule (how he wills to act). (*MM*, 6:225)[3]

In each of [M1]-[M3] Kant describes maxims as subjective practical principles or subjective principles of action. They are *subjective* both because a subject prescribes them to herself (Kant writes that they are "rules imposed upon oneself" (*G*, 4:438)) and because they reflect her own circumstances and point of view (often, maxims are adopted just because of a person's inclinations or because of her ignorance of things). They are *practical* because they guide a person's behavior (as it is put in [M1], a maxim is a "principle in accordance with which the subject *acts*"). Finally, they are *principles* because of their generality; maxims apply to situation-*types*. As Kant puts it in [M2], maxims are "propositions that contain a *general* determination of the will, having under {that determination} several practical rules".

These three marks of maxims—their subjectivity, practicality, and generality—apply both to policy- and disposition-maxims. Consequently, although Kant's definitions of maxims give the appearance that he understands the term to apply to only one kind of thing, one need not read them that way. Instead, the definitions are broad enough to cover both policy- and disposition-maxims, so both kinds of maxims equally deserve the appellation, "maxim".

I shall explicate Kant's conception of maxims by answering several common questions about them. First, how often do agents act on maxims? Second, what is the degree of generality of the situations to which maxims apply? And third, how aware are

agents of their own maxims? My answers to these questions are as follows. First, agents cannot act without acting on a maxim. Second, a maxim can apply to a narrowly circumscribed situation-type, to a kind of situation that an agent may regularly encounter, and even to a situation-type so general that she faces it all the time. Third, an agent's awareness of her maxims can vary greatly, usually depending on the scope of the maxim; she might know precisely what her maxim is, or she might not know at all the maxim on which she acts. In general, the greater a maxim's generality, the less transparent that maxim will be to the person. I shall now provide evidence for each of these conclusions.

2.1. The Maxim-Action Thesis

While Kant only once says that agents always act on maxims ("as a freely acting being, a human being actually cannot do anything without the will – he acts always according to maxims even if not universally" (*LM-D*, 28:678)), we can find a clear commitment to the thesis that agents always act on maxims—which I shall call the "Maxim-Action Thesis"—once we realize that every maxim is associated with what Kant calls a *Triebfeder*, which is usually translated as "incentive".[4]

In *Religion within the Boundaries of Mere Reason*, Kant advances what has come to be known (due to Henry Allison[5]) as the "Incorporation Thesis" (IT): "freedom of the power of choice has the characteristic, entirely peculiar to it, that it cannot be determined to action through any incentive *except so far as the human being has incorporated it into his maxim* (has made it into a universal rule for himself, according to which he wills to conduct himself)" (*Rel*, 6:23-4). According to the IT, a person is never "determined to action" by an incentive unless she first "incorporates" it into a maxim. In other words,

whenever a person acts on an incentive, she also acts on a maxim in which that incentive plays some part.

Kant thinks that people cannot act without acting on an incentive; call this the "Incentive-Action Thesis". In student lecture notes taken between 1782 and 1783, Kant says that "should I desire neither according to understanding nor to sensibility, then I would want that which displeases me, I would act without incentive and cause, and that is impossible … we cannot be determined by mere representations of reason; it must also give us incentives" (*LM-M*, 29:900) and "we do always act … according to the larger multitude of incentives and thus not according to mere whim" (*LM-M*, 29:902). Assuming these lecture notes accurately convey Kant's thinking,[6] we can make the following argument:

(1) A person cannot act without acting on an incentive (the Incentive-Action Thesis);

(2) for a person to act on an incentive, she has to incorporate it into a maxim (the IT);

(3) therefore, a person cannot act without acting on a maxim (the Maxim-Action Thesis).

There remains room to be skeptical of the Maxim-Action Thesis, for I have offered only two passages that substantiate one of its supports, viz., the Incentive-Action Thesis, and those passages both appear approximately ten years before Kant's articulation of the IT; thus, while the Maxim-Action Thesis requires both the IT and the Incentive-Action Thesis to be true, it is not clear that Kant even held the Incentive-Action Thesis, and it is further not proven that he held both theses simultaneously.[7] However, there are many other passages, spanning many years, that show that Kant accepted the Incentive-Action Thesis, so we can be sure that Kant maintained the Maxim-Action Thesis at least

by the time he advanced the IT. But I think we can go further: while in no passage before 1793 does Kant directly assert the IT, on numerous occasions he advances the claim that incentives alone cannot determine us to action, unless they have the help of reason. While this is not identical to the IT, it is similar enough for us to conclude, I think, that he held something that for all practical purposes is the same.

To appreciate the passages I am about to mention, one has to first of all grasp Kant's concept of the *Willkür*. The *Willkür* is what allows a person *to try* to act on her inclinations; if a person had inclinations but no *Willkür*, she would be unable *to attempt* to act on them.[8]

Kant lists three kinds of *Willkür*—free, affected, and animal:

> That choice which can be determined by *pure reason* is called free choice. That which can be determined only by *inclination* (sensible impulse, *stimulus*) would be animal choice (*arbitrium brutum*). Human choice, however, is a choice that can indeed be *affected* but not *determined* by impulses, and is therefore of itself (apart from an acquired proficiency of reason) not pure but can still be determined to actions by pure will. (*MM*, 6:213)

A being with a *Willkür* that can be determined to try to act by pure reason alone, without being affected by a sensible impulse or stimulus, has a free *Willkür*. A *Willkür* that is determined by impulse or stimulus alone is an animal *Willkür*. However, people have affected *Willküren*; a human *Willkür* is *always* affected by impulses or stimuli, but is only determined to attempt to act by reason.[9]

Kant's claim that human *Willküren* are only *affected* by impulse or stimulus, but *determined* by reason, is not something he formulates only in 1797. Rather, Kant attributes to us an affected *Willkür* throughout his critical period. For instance, in 1784 he says, "Human choice is an *arbitrium liberum* {free choice}, in that it is not necessitated

per stimulus" (*LE-C*, 27:267), and in 1792-93 he claims, "We are *affected* by stimuli, but *not determined*" (*LM-D*, 28:677).[10]

Admittedly, the above passages state that human *Willküren* are free from determination by *stimuli*, not *incentives*; however, incentives and stimuli can be, and usually are, the same thing: "That which is the cause of the desires is the impelling cause or incentive of the soul. Now, if they arose from sensibility then they are called stimuli and their effect desire aroused by stimuli or sensible desires" (*LM-M*, 29:895).[11] Desires are aroused by "impelling causes" or incentives; if that incentive is sensible rather than intellectual (i.e., if it comes from something besides one's pure practical reason), then it is a stimulus. Thus, in saying that the human *Willkür* is not determined by stimuli, Kant is saying that our sensible incentives never make us act without reason also playing a role.

Thus, from the 1780s to the 1790s Kant holds to the thesis that the human *Willkür* is one that is *affected* by sensible incentives, but not *determined* by them. The fact that people, whenever they act, are always affected by some incentive (sensible or intellectual),[12] proves the Incentive-Action Thesis, and the fact that people are determined to action only through their reason, supports, though does not prove, the Maxim-Action Thesis (as we shall see, reason "determines action" through motives, not maxims; but incentives and motives together constitute maxims). Nonetheless, I shall take the Maxim-Action Thesis to be provisionally established; section 2.2 provides more support.

2.2. What Maxims Are

There is a significant problem for the Maxim-Action Thesis, however. According to that thesis, people always act on maxims whenever they act, but according to his

theories of education and character, Kant is clear that people do not always act on maxims. Thus, the Maxim-Action Thesis appears to conflict with Kant's theories of education and character. In this section, I dissolve the appearance of conflict by distinguishing between two kinds of maxims that can be found in Kant's corpus: policy-maxims, which are the kind of maxims Kant discusses in his theory of education and character, and disposition-maxims, which are the kind of maxims Kant usually discusses in his ethical works. My claim, then, is that the Maxim-Action Thesis holds that people always act on *disposition*-maxims, while Kant's doctrines of education and character require people to learn how to act on *policy*-maxims.

2.2.1. Policy-Maxims

In *Education*, Kant emphasizes that children need to learn two things to become useful members of society: discipline, which is the ability to resist impulsive action on one's inclinations, and character, which is the ability to act according to maxims:

> *Moral culture* must be based upon 'maxims,' not discipline; the one prevents evil habits, the other trains the mind to think. We must see, then, that the child should accustom himself to act in accordance with 'maxims,' and not from certain ever-changing springs of action. … The child should learn to act according to 'maxims,' the reasonableness of which he is able to see for himself. (*E*, §77)

If children have to learn to act according to maxims, then clearly there is a time at which they do not know how to act on maxims.[13]

Kant also writes in *Education* that "Character consists in readiness to act in accordance with "maxims"" (*E*, §78). Because Kant believes that "The first endeavor in moral education is the formation of character" (*E*, §78) and educators' "ultimate aim is the formation of *character*" (*E*, §94), it follows that it is important to teach people to act on maxims because that skill is needed to have character, and character is good. But why is character good?

Without character a person will be a slave to her inclinations: "if a man be allowed to follow his own will in his youth, without opposition, a certain lawlessness will cling to him throughout his life" (*E*, §5).[14] As Kant puts it in *Anthropology from a Pragmatic Point of View*, a person without character will "fly off hither and yon, like a swarm of gnats" (*Ant*, 7:292).[15] Consequently, since one needs to be able to act consistently (which requires constant vigilance against one's own inclinations[16]) to live in conformity to the moral law, character—readiness to act in accordance with self-imposed maxims—is necessary to live a morally good life; that is why character is good.[17]

Unfortunately, character is rare: "To be able to simply say of a human being: 'he has a *character*' is not only to have *said* a great deal about him, but is also to have *praised* him a great deal; for this is a rarity" (*Ant*, 7:291-92). Given that people with character are people who firmly guide themselves by their own maxims, and given as well that such people are rare, it follows that most people do not usually firmly adhere to their own maxims.

According to Kant's conception of character, most people, most of the time, do not act on maxims. But according to the Maxim-Action Thesis, people always act according to maxims. It appears as though Kant is contradicting himself.

If we look more closely at the maxims he thinks people ought to act on, though, we can see that the appearance of self-contradiction is illusory. In the *Anthropology*, Kant recommends some "principles that relate to character" (which I take to be maxims), such as "Not intentionally to say what is false"; "Not to break one's (legitimate) promise"; and "Not to pay attention to gossip derived from the shallow and malicious judgment of others" (*Ant*, 7:294).

Other examples of maxims can be found in the *Critique of the Power of Judgment*, where Kant discusses the "maxims of the common human understanding": "think for oneself"; "think in the position of everyone else"; and "think in accord with oneself" (*CJ*, 5:294). These maxims, as well as the ones listed in the *Anthropology*, are clearly maxims by which a person could easily fail to abide. It is not surprising that it would take significant attention from educators, as well as dedication on the part of individuals, to produce someone who firmly adhered to maxims such as those.

But I do not believe that those maxims are the kind of maxims on which people always act. Instead, I take those maxims to be policy-maxims, for they are self-imposed policies that apply to a wide variety of situations.[18] These could not be the kind of maxims on which all people, almost all the time, act, for doing so requires a lot of diligence. Instead, I take the maxims on which all people, almost all of the time, act to be disposition-maxims.

2.2.2. Disposition-Maxims

Before describing disposition-maxims in greater detail, it is worth giving a word about why I call them "disposition-maxims" in the first place. The reason is textual: Kant often uses "disposition" interchangeably with "maxim". For example, "Ethics is … a philosophy of dispositions, and hence a practical philosophy, for dispositions are basic principles of our actions and serve to couple actions with their motivating ground" (*LE-C*, 27:299) and "Of every action that conforms to the law but is not done for the sake of the law, one can say that it is morally good only in accordance with the *letter* but not the *spirit* (the disposition)" (*CPrR*, 5:72).[19]

Similarly, in the *Religion* Kant sometimes equates a person's supreme disposition (i.e., her *Gesinnung*) with a maxim or ground of actions. For example, "The *Gesinnung*, i.e. the first subjective ground of the adoption of the maxims, can only be a single one, and it applies to the entire use of freedom universally" (*Rel*, 6:25), and "the transformation of the *Gesinnung* of an evil human being into the *Gesinnung* of a good human being is to be posited in the change of the supreme inner ground of the adoption of all the human being's maxims in accordance with the ethical law" (*Rel*, 6:51).[20]

Most important, for present purposes, about Kant's use of "disposition" or "*Gesinnung*" is his recognition of the fact that a person's disposition(s) or *Gesinnung* is hard to know. For instance, "The disposition cannot be required by the sovereign, since it is not known, in that it is internal" (*LE-C*, 27:273).[21] Similarly, "we can draw inferences regarding the *Gesinnung* only on the basis of actions (which are its appearances)" (*Rel*, 6:70n).[22] In addition, the term "disposition" itself indicates that Kant means it to refer foremost to the way a person *tends* or is *disposed* to act in various circumstances; and there no reason for thinking that a person should be able to know with any certainty how either she or other people tends to act in particular situation-types.

Thus, I conclude that when Kant uses "disposition" interchangeably with "maxim", he means by it to refer to a kind of maxim that is hard to discover, either in oneself or others.[23] Given, first, the Maxim-Action Thesis, and second, the plausible claim that people are often confused as to why they do the things they do, I conclude that the kinds of maxims on which people always act (i.e., the kind of maxim that is the subject of the Maxim-Action Thesis) are dispositions. Hence, I call such maxims "disposition-maxims."[24]

2.2.3. The Elements of Disposition-Maxims

Kant has a complex account of the elements that compose disposition-maxims. A disposition-maxim always contains two elements: an incentive (*Triebfeder*)/stimulus (*Anreiz*, *Begwegungsursache*, or *Reiz*) and a motive (*Bewegungsgrund* or *Motive*). Elucidating these elements not only clarifies what a disposition-maxim is, but also reveals Kant to be a hedonist about non-moral motivation and value and also shows the relationship between imperatives and maxims.

The story I shall tell is this: every action results from a combination of incentives and motives (and because every action is action on a disposition-maxim, disposition-maxims are combinations of incentives and motives). An incentive makes a person want to bring about a state of affairs,[25] while a motive makes her see some action state of affairs as good, that is, as something that should be undertaken or brought about.[26] Incentives are psychological states that motivate on the basis of expected pleasure, while motives are endorsements of imperatives that result in judgments of goodness.

The causal relationship between an incentive and a motive crucially determines the moral worth of an action. If someone experiences an incentive and *because of the incentive* judges that a certain action or state of affairs is good—if, that is, the incentive produces the motive—, then that person's action has no moral worth. On the other hand, if someone judges that some action or state of affairs is good, and *because of that judgment* feels motivated to undertake the action or bring about that state of affairs—if, that is, the motive produces the incentive—, then the action has moral worth. Kant calls incentives that produce motives "sensible" incentives or "stimuli", and the motives they cause "pragmatic" motives; a motive that produces an incentive he calls a "practical" or

"moral" motive and its corresponding incentive an "intellectual incentive", "moral feeling", or "respect".

There is not much explicit textual evidence for the claim that Kant takes all action to spring from a combination of incentives and motives, though there is this passage:

> the *concept of freedom* rests on this: *namely the faculty of a human being for determining oneself to action through motives, independently of the sensible impulses affecting him,* therefore the power of choice is free only insofar as it is not determined by stimuli … {the human being's} power of choice is thus in part sensual, in part intellectual, where the effecting stimuli are indeed to be viewed also as a ground for action, but only as an insufficient ground, the motives, on the other hand, contain that ground of determination which makes the action necessary. (*LM-K*$_3$, 29:1015-16)

There are many passages where Kant, in adumbrating the nature of action, lists both incentives and motives as the only kinds of causes of action,[27] but unfortunately, only the quotation above explicitly asserts that *both* motives and incentives are necessary for the production of action.[28] Still, I hope to make clear in what follows that the best interpretation of what Kant believed, *throughout his critical period*, is that an incentive and a motive are both necessary ingredients of any individual action.

2.2.3.1. Incentives

At this point it is necessary to define incentives and motives. Kant tends to define the two together. The following two passages, the first from 1782-1783, and the second from 1794-1795, are representative:

> That which is the cause of the desires is the impelling cause or incentive of the soul. Now, if they arose from sensibility then they are called stimuli and their effect desire aroused by stimuli or sensible desires. But if they originate from the understanding, then they are called motives, their effect [is called] desire aroused by motives or intellectual desires. (*LM-M*, 29:895)

> There lie in human beings … incentives of the soul or grounds of determination, sources of the possibility for producing the represented, determining or impelling causes, and these *lie either in the understanding as in the law of action, or in the sensibility, namely, in the feeling of pleasure and displeasure,* and are therefore either sensitive causes and incentives or intellectual causes and incentives. Thus arises the division of the concept into *the higher or rational power of choice*, i.e. the faculty of desiring through motives, or will or the power of choosing from an impelling intellectual cause, and sensitive power of choice, the faculty of desiring through stimuli. (*LM-K*$_3$, 29:1014-15)[29]

An incentive is a mental state that inclines a person to action or, what is the same thing, can give rise in her to a desire for something: "by *incentive* (*elater animi*) is understood the subjective determining ground of the will of a being whose reason does not by its nature necessarily conform with the objective law" (*CPrR*, 5:72). Notice that incentives cause desires—thus, the two are not identical.

What, then, is a desire, and how does it relate to an incentive? Kant defines a desire as follows: "*Desire* … is the self-determination of a subject's power through the representation of something in the future as an effect of this representation" (*Ant*, 7:251). He goes on to compare desires to wishes: "Desiring without exercising power to produce the object is *wish*" (*Ant*, 7:251). These two definitions suggest that you desire the reality of some representation only when you try to actualize it;[30] if you would like something to be the case, but do not try to bring it about because you think it cannot be done, then you merely wish for it. So trying to bring about a state of affairs is the same thing as desiring it.

This is not how we customarily use the term "desire". We often speak of people who have desires without trying to act on them, or people who have conflicting desires. Yet on Kant's way of speaking, neither of these things is possible. To avoid any misunderstanding, then, I shall refer to desiring something in the Kantian sense—that is, trying to bring about a particular state of affairs—as "K-desiring", and desiring in the normal sense simply as "desiring".

Nevertheless, if we understand desires as Kant does—if, that is, we take them to be K-desires—then it results in an understanding of incentives that makes them turn out to be desires (in our sense of the term). Incentives, after all, are the subjective

determining grounds of K-desires ("The subjective ground of desire is an *incentive*" (*G*, 4:427)). That is, a necessary condition for a person forming a K-desire for *X* (i.e., for trying to realize state of affairs *X*) is that *X* provides her with an incentive (i.e., that she feels subjectively motivated to bring about *X*); in other words, incentives are psychological states, the possession of which inclines a person to try to bring about a certain state of affairs. And this just makes incentives seem like desires as we understand them. On this understanding of incentives, then, the Incentive-Action Thesis is really the Desire-Action Thesis; it amounts to the claim that you cannot try to act without having a desire for something—a claim that will no doubt strike many people as common sense. Moreover, Kant has a way of accounting for the possibility of having a desire but not acting on it—this is just having an incentive but not acting on it. Similarly, Kant can account for conflicts in desires—this just amounts to having competing incentives.

Still, there seem to be a couple of wrinkles in interpreting incentives as desires (in our sense) and K-desires as attempted actions. First, Kant defines inclinations as "Habitual sensible desire{s}" (*Ant*, 7:251).[31] Yet in a variety of places, he says that pure practical reason can resist inclination;[32] e.g., "even when {people} do obey the law, they do it *reluctantly* (in the face of opposition from their inclinations)" (*MM*, 6:379). Thus, even though a K-desire is an attempted action, having an inclination—a habitual, sensible K-desire, that is, one we often engage in—apparently does not imply that the agent whose inclination it is has attempted to perform any action. This is strange: if the only difference between a K-desire and an inclination is that an inclination is a K-desire that a person often engages in, should not having an inclination result in an attempted action just as surely as having a K-desire does?

No. Because inclinations are *habitual* K-desires, it follows that particular inclinations—e.g., the sexual inclination—represent *drives*, or *sources of types of incentives*.[33] For instance, if you know that someone has an inclination to game-playing, you know that she often has K-desires for playing games—i.e., she regularly tries to play games. Presumably, though, the explanation for why she regularly tries to play games is that she is constituted such that she finds games appealing. And this is just to say that games in general provide her with incentives. So to say that someone can resist her inclination to game-playing is not to say that she has K-desires and somehow overcomes them; rather, it is to say that she can resist the incentives the games provide her, thereby refraining from acting on her drive for game-playing.

The other problem confronting the person who interprets incentives as desires and K-desires as attempted actions is distinguishing actions from K-desires. It seems to me that there is no good solution to this problem. But here is a speculation (unsupported by any text): the difference between K-desires and actions is that actions are a subset of K-desires: whereas K-desires are attempts to carry out desires, an action is a *successful* K-desire, i.e., not only an attempt to act on a desire, but an attempt that actually succeeds. An unsuccessful K-desire is not a wish—wishes are not K-desires at all, but rather incentives accompanied with the belief that it is impossible to realize the state of affairs responsible for the incentive—but rather an attempt to act on a desire where it is impossible to act on that desire. In other words, a K-desire that is not an action is a K-desire held by an agent who has the false belief (or beliefs) that she can do something that she, in fact, cannot do.

Incentives, then, are garden-variety desires. But we need to say something else about incentives: a representation of an object can provide an incentive only if the agent believes that its realization will bring about pleasure. Kant writes, "incentives of the soul or grounds of determination … *lie either in the understanding as in the law of action, or in the sensibility, namely, in the feeling of pleasure and displeasure*" (*LM-K*$_3$, 29:1014). Incentives are therefore psychological states that motivate us to action through the expectation of pleasure. In our language, on Kant's view we desire things only if we think those things will give us pleasure.[34]

Given that we cannot K-desire anything without having an incentive that is its basis, it follows that we can K-desire something only if we think it will give us pleasure: "Each desire is grounded in the sense of anticipated pleasure" (*LM-L*$_2$, 28:587).[35] This applies to all K-desires—actions undertaken to satisfy a sensible incentive, and action done out of respect for the moral law:

> since we desire merely that which pleases us, pleasure is the cause of our desiring. But the cause of the pleasure is either sensibility or understanding. No third thing is possible. Thus if I do not desire something from sensibility, then I must desire it because it pleases my understanding. (*LM-M*, 29:900)[36]

How can it be that even when we act out of respect for the moral law, we are acting for pleasure? Answering this question requires a long excursion into motives, the second element of disposition-maxims.

2.2.3.2. Motives

Someone subject to an incentive wants to bring about a particular state of affairs, but Kant is insistent that people are never (except when controlled by affects) determined to action by incentives. Instead, people are only *affected* by incentives; they require a motive to act: "stimuli can never determine {the *Willkür*}, but rather merely affect it

sensibly, and in order to determine it there remains necessary the concurrence of the understanding" (*LM-K*$_3$, 29, 1015).[37] Incentives by themselves only incline a person to act; in order for her to move from wanting to acting, a person needs the judgment that what she desires is good, and this judgment is contributed by the motive: "*good* or *evil* always signifies a reference to the *will* insofar as it is determined by the *law of reason* to make something its object; for, it is never determined directly by the object and the representation of it, but is instead a faculty of making a rule of reason the motive of an action (by which an object can become real)" (*CPrR*, 5:60).[38]

My claim that motives contribute judgments of goodness is not something that is easily extracted from the texts; one reason for this is that Kant sometimes gives the impression that moral motives are the only kind of motive there is. For instance, in the *Groundwork* he writes, "The subjective ground of desire is an *incentive;* the objective ground of volition is a *motive;* hence the distinction between subjective ends, which rest on incentives, and objective ends, which depend on motives, which hold for every rational being" (*G*, 4:427).[39] Here it looks as though we act from motives only when we act to promote objective ends, that is, ends we are commanded to have by the moral law.

But Kant clearly believes that there are also pragmatic motives in addition to moral motives: "the authors of that {Wolffian} science remain true to their idea of it in this too; they do not distinguish motives that, as such, are represented completely a priori by reason alone and are properly moral from empirical motives, which the understanding raises to universal concepts merely by comparing experiences" (*G*, 4:391).[40]

There are thus both pragmatic and moral motives. But it is still not clear what a motive *is*. Perhaps Kant's clearest definition runs: "We first have to take up two points

here: (1) The principle of appraisal of obligation, and (2) the principle of its performance or execution. Guideline and motive have here to be distinguished. The guideline is the principle of appraisal, and the motive that of carrying-out the obligation" (*LE-C*, 27:274). From this definition, it appears that motives are responses to—more exactly, "carryings-out" of—principles of obligation, or imperatives. In other words, a motive is an agent's endorsement of an imperative.

To prevent confusion, I need to say one thing about my term, "endorsement". The endorsement of an imperative could refer to two different things in this context. On the one hand, I could endorse an imperative if I believed that it was true. For instance, I might endorse the imperative, "if you want to poison someone, you ought to use cyanide" simply by believing that it accurately describes a causal relation ("*I* don't want to poison anyone, but if I did, I agree that cyanide would be an effective means"). On the other hand, I could endorse an imperative not just by accepting that the causal relations it assumes really hold, but by *applying* that imperative to my own case and as a result finding it particularly relevant. For example, if I wanted to poison someone, then that hypothetical imperative would take on a different complexion; it would no longer be a merely theoretical statement, but one with a newfound significance. When I talk about motives as being endorsements of imperatives, I mean "endorsement" in this second sense of "making practically relevant."[41]

A pragmatic motive differs from a moral motive in that a pragmatic motive is an endorsement of a particular hypothetical imperative whereas a moral motive is an endorsement of a particular categorical imperative. Endorsements do not take place in conscious moments of decision; rather, we know that someone has endorsed an

imperative when she sees what she is obligated to do (according to that imperative) as good or obligatory: "All imperatives are expressed by an *ought* and indicate by this the relation of an objective law of reason to a will that by its subjective constitution is not necessarily determined by it (a necessitation). They say that to do or omit something would be good" (*G*, 4:413).[42] In other words, we know that a person endorses an imperative—i.e., has a motive—when she judges some action to be obligatory or good (or some state of affairs to be one she is obligated to bring about or to be good to realize). Motives, then, always bring with them judgments of goodness.

Someone who endorses the hypothetical imperative "if you want popularity, then you ought to flatter others",[43] sees flattering others as good; however, she endorses the hypothetical imperative in the first place because personal popularity is an incentive to her—she wants to be popular because she thinks it would please her, so she judges as good those steps she needs to take to become popular: "if {an} action would be good merely as a means *to something else* the imperative is *hypothetical*" (*G*, 4:414).

Categorical imperatives differ from hypothetical imperatives in that a person cannot refrain from endorsing them: "it is never stimuli alone that determine me to the action: a representation, even if unclear, of the law of duty is always concurring alongside" (*LM-K*$_3$, 29:1015). Because the Categorical Imperative is always operative in a person[44] through her conscience,[45] she sees immoral actions as evil and morally required actions as good:[46] "if the action is represented as *in itself* good, hence as necessary in a will in itself conforming to reason, as its principle, *then it is categorical*" (*G*, 4:414).

Note that there are two kinds of goodness—moral goodness and prudential goodness. When you endorse a categorical imperative, you see an action as good because

it is commanded by the moral law, whereas when you endorse a hypothetical imperative you see an action as good because it is itself pleasing to engage in, or because it helps to produce a state of affairs that would be pleasing.[47]

Given what we have established about motives, we can make the following argument:

(4) To act, a person needs to have a motive;

(5) a person has a motive when she endorses an imperative;

(6) an agent who endorses an imperative judges some action to be obligatory, or what is the same thing, good to do; therefore,

(7) to act, a person needs to judge some action to be obligatory, or what is the same thing, good to do.

According to this argument, Kant is committed to the view that a person cannot act without judging that action to be good to do. However, Kant at times seems to indicate that the only actions we judge as good are those commanded by the moral law: "What we are to call good must be an object of the faculty of desire in the judgment of every reasonable human being, and evil an object of aversion in the eyes of everyone; hence for this appraisal reason is needed, in addition to sense" (*CPrR*, 5:61).

His statements to this effect are often ambiguous, though. For instance, Kant writes in the *Groundwork* that "Practical good … is that which determines the will by means of representations of reason, hence not by subjective causes but objectively, that is, from grounds that are valid for every rational being as such" (*G*, 4:413); note that Kant here speaks only of "practical" good, not goodness generally. Similarly, in his lectures on ethics Kant says that the "good must be distinguished from the pleasant; the pleasant

refers to sensibility, the good to the understanding. The concept of the good is a thing that satisfies everyone, and hence it can be judged by the understanding" (*LE-C*, 27:264). Shortly after making that statement, though, he adds that "the good can be good in a variety of ways for any given purpose, since it is a principle of skill and prudence; only if it is good for moral actions would it then be a moral principle" (*LE-C*, 27:264).

This suggests that Kant's statement in the second *Critique* refers just to the concept of practical goodness, not goodness in general. Given Kant's other statements about goodness—e.g., "imperatives express objective necessitation, and since they are of three kinds {problematic and pragmatic (i.e., hypothetical), and moral (i.e., categorical)}, there are also three kinds of goodness" (*LE-C*, 27:255)[48]—we have strong reason to think that Kant endorsed (7) (to act, a person needs to judge her action to be good).

2.2.3.3. Intellectual Incentives and Pleasure[49]

Whereas a person's prior experience of a sensible incentive moves her to endorse a hypothetical imperative (which endorsement amounts to a pragmatic motive), a person's endorsement of a categorical imperative (which endorsement amounts to a moral motive) produces an intellectual incentive, respect for the moral law: "respect for the moral law is a feeling that is produced by an intellectual ground" (*CPrR*, 5:73).

A peculiar feature of respect is that it cannot exist without first of all conflicting with a desire. That is, before a person can feel respect for the moral law, it first of all has to "thwart" an immoral desire:

> What is essential in every determination of the will by the moral law is that, as a free will … it is determined solely by the law. So far, then, the effect of the moral law as incentive is only negative, and as such this incentive can be cognized a priori. For, all inclination and every sensible impulse is based on feeling, and the negative effect on feeling (by the infringement upon the inclinations that takes place) is itself feeling. Hence we can see a priori that the moral law, as the determining ground of the will, must by thwarting all our inclinations produce a feeling that can be called pain (*CPrR*, 5:72-73).

Kant starts with the premise that "What is essential to any moral worth of actions is *that the moral law determine the will immediately*" (*CPrR*, 5:71). That is, an action does not have moral worth unless it is done, not only in conformity with the moral law, but *because* it is in conformity with the moral law; if, for instance, someone fulfils a duty, but only because she wanted pleasure, then it is a matter of luck that she avoided immorality (i.e., if fulfilling the duty would not have given her pleasure, she would not have done so), and so her action has no moral worth.

Consequently, for someone to do something *because* it is conformity with duty, she has to resist her inclinations, either because they propose actions that directly transgress the moral law, or because they tempt her to rely on them as the primary bases for her actions.[50] In either case, whenever a person recognizes that duty to morality is her fundamental obligation, she is forced to oppose an inclination within herself, and this feeling of conflict causes her pain. But her recognition of morality as her fundamental obligation also dampens the strength of her inclinations. The fact that a mere recognition that something is her duty can weaken the force with which her inclinations assail her arouses in her a sort of awe at her own practical reason, which Kant calls respect. Thus, respect results from recognizing that something is your duty, and this respect can itself serve as an incentive—an intellectual incentive—motivating you to do your duty.[51] Hence Kant's claim that "though respect is a feeling, it is not one *received* by means of influence; it is, instead, a feeling *self-wrought* by means of a rational concept" (*G*, 4:401n).[52]

How does respect serve as an incentive? Incentives, after all, are desires to perform actions or bring about states of affairs because of their pleasure. Like sensible

incentives, respect is also a desire to undertake an action or bring about a state of affairs; unlike them, though, respect is not the desire to do something *because* of pleasure – rather, respect *is* a kind of pleasure,[53] one that gives rise to a desire to do something.[54] It works like this: when your representation of the moral law infringes on an inclination, you feel intellectual pain for inappropriately having the inclination in the first place, but also pleasure from being able to weaken your inclinations simply for being inappropriate. This pleasure is respect—you *respect* the moral law in yourself for its power to still another part of yourself, your natural desires. *Because* you are impressed by the moral law's power (which impression is a kind of pleasure), you carry out its directives (this is not mere slavishness to authority, though, because the moral law's directives represent what you are rationally committed to, i.e., what you truly want); this is how we can act *from respect*. Thus, when you act from respect, you do not act *to experience* pleasure, but rather *for the sake of* pleasure you experience (because you feel impressed, you act in a certain way, and if you are successful in carrying out the action, you will feel even more pleasure, a pleasure Kant calls "self-satisfaction" or "self-contentment").[55]

2.2.4. Combining the Elements into Disposition-Maxims[56]

Let me summarize what I have shown so far. Incentives are desires to undertake actions or bring about states of affairs, either for pleasure (in which case the incentive is sensible) or for the sake of pleasure (in which case the incentive is intellectual). Motives are endorsements of imperatives, which endorsements constitute judgments that an action or states of affairs is good. According to the Maxim-Action Thesis, all action is action on a disposition-maxim; but according to what I have shown in section 2.2.2.1, all action is

action on both an incentive and a motive. Thus, I take it that disposition-maxims are comprised by incentives and motives.

Another reason for thinking that incentives and motives compose disposition-maxims comes from the fact that actions performed from moral motives have moral worth: "one who does a thing that is good in and for itself, is acting from motives" (*LE-C*, 27:257).[57] However, an action can have moral worth only if it is done from respect for the law: "when moral worth is at issue, what counts is not actions, which one sees, but those inner principles of actions that one does not see" (*G*, 4:407). Thus, another reason for thinking that disposition-maxims contain motives as elements is that Kant sometimes writes that an action can have moral worth only if it has the right kind of motive, and at other times writes that an action can have moral worth only if it has the right kind of maxim.

A final reason for thinking disposition-maxims to be composed by incentives and motives comes from a passage from Vigilantius's notes on Kant's lectures on ethics:

> The *maxim* of an action … includes all subjective grounds of action whatsoever, insofar as they are taken to be real. … the maxim is the subjectively practical principle, insofar as the subject makes the rule by which he is to act into the motive of his action as well. It is the maximum in determination of the grounds of action. (*LE-V*, 27:495)[58]

This suggests that every force that helps to bring about action—not only motives, but also incentives—gets included in maxims.

Whenever anyone acts on a disposition-maxim, she acts on an incentive and a motive (given the Maxim-Action Thesis, whenever anyone acts, period, she acts on an incentive and a motive). She judges that some action or state of affairs is good (i.e., she thinks she ought to perform some action or realize some state of affairs), and also desires

to undertake that action or bring about that state of affairs. In judging something good, though, she accepts a general principle that, in certain circumstances, things of a certain sort are good. That is, she accepts the principle, "in circumstances C, action A is good"[59] or "in C, I ought to do A".

Merely accepting a principle, though, is not the same thing as having a disposition-maxim; it is just making a judgment (having a motive). Similarly, if you only have the desire to do A, you do not have a disposition-maxim; you merely have a desire (an incentive).[60] To have a disposition-maxim, you need not only to *judge* that A is good, you also have to *desire* to do A. Thus, a person has a disposition-maxim when she both wants to do something and judges that it is good—a disposition-maxim is thus a *motivating* judgment that some action in some circumstances is good. (The incorporation of an incentive into a maxim, then, is the coupling of a desire for an object with a judgment of that object's goodness.)

A person can experience several disposition-maxims at the same time: she can judge several different things to be good, and have desires for them all. However, when she acts, she selects one of these disposition-maxims to act on. In Kant's language, a person makes a desire the determining ground of her will (i.e., has a K-desire) through a maxim: "It is … strange that intelligent men could have thought of passing off the desire for happiness as a universal *practical law* on the ground that the desire, and so too the *maxim* by which each makes this desire the determining ground of his will, is universal" (*CPrR*, 5:28).

3. The Flexible Maxim Thesis

We have answered the first question about maxims I posed above, viz., "how often do agents act on maxims?" According to the Maxim-Action Thesis, agents act on maxims—disposition-maxims—every time they act. We are now positioned to answer the second question, "what is the degree of generality of the situations to which maxims apply?" I answer that disposition-maxims can apply to situations of any degree of generality (I call this the "Flexible Maxim Thesis"), while policy-maxims apply only to situation-types that are at least fairly general.

Take policy-maxims first. Kant's examples of policy-maxims—e.g., "Not intentionally to say what is false"—show them to be fairly general. The policy-maxim, "don't lie" applies to any situation in which lying is an option: perhaps whenever you communicate; alternatively, any situation in which you have something to gain by concealing your intentions. Given its generality, it requires a high degree of practical judgment to apply well; as Kant's "casuistical questions" about lying in *The Metaphysics of Morals* show,[61] lying is (perhaps) blameworthy only when the target of your lie expects truthfulness from you (for example, most of the time when someone asks you how you are doing, she does not in fact want a report of your long-standing psychological problems).

On the other hand, policy-maxims will be useless if too specific ("if you're writing page 103 of your novel, make sure not to use the word, 'aircraft'") or too general ("don't do anything wrong"). Thus, policy-maxims are not supposed to apply to situations that are either too narrow or too broad.

The situation is much different with disposition-maxims. As was made clear in sections 2.1 and 2.2.1, every action is undertaken on a particular disposition-maxim. A

disposition-maxim, though, includes a judgment of something's goodness and the desire for something's realization. Consequently, the kinds of situation a disposition-maxim applies to depend on what is desired and what is judged good. Since extremely general or specific outcomes can be desired or judged good, it follows that disposition-maxims can apply to extremely general or extremely specific situations.

Here are some examples. Suppose a person is hungry and sees a tasty sandwich. Not only will he want to eat the sandwich, he will see eating the sandwich as good—i.e., he will see that course of action as providing him with at least a *prima facie* reason for action. That is, he will accept the disposition-maxim, "when I am hungry, I should eat what will satisfy my hunger" or "when I am hungry, satisfying my hunger is good." If he actually eats the sandwich, then he *acts on* that disposition-maxim.

Imagine, however, that he is on a diet, and the sandwich is highly caloric. In such a case, he still accepts the disposition-maxim, "when I am hungry, I should eat what will satisfy my hunger", because, after all, he sees eating the sandwich as a good thing to do, as having something going for it. However, he also accepts the disposition-maxim, "when I am on a diet, I should not eat highly caloric things" or "when I am on a diet, eating highly caloric things is bad". As a result, eating the sandwich appears to him to be both good and bad; good, because it will satisfy his hunger, but bad, because it will undercut his diet. In having this mixed reaction, he shows himself to accept two conflicting disposition-maxims. Which one we should say he acts on, though, depends on what he ends up doing; obviously, if he ends up eating the sandwich, he acts on the sandwich-maxim, but if he ends up refraining, he acts on the diet-maxim. In either case, though, he still accepts both maxims—if he did not, deciding whether or not to eat the sandwich

would be easy. In fact, though, both courses of action to present to him what he takes to be *pro tanto* practical reasons.

Both the sandwich- and the diet-maxim are fairly general—they apply to situations of a type a person regularly encounters, but neither applies to, say, a one-time or extremely delimited event, or to every kind of situation that a person encounters. However, disposition-maxims of both sorts can be imagined.

Pretend God appears to you and tells you that you should help Hildreth Milton Flitcraft with her model train hobby by giving her exactly $11.15 this Tuesday at exactly 2:47 pm.[62] When the appointed moment arrives, you are in a position to give Hildreth exactly $11.15. Since you accept that God has a good reason for rendering this command, you see giving exactly $11.15 to Hildreth this Thursday at exactly 2:47 as having a lot going for it. This is true even though you do not care one whit for: the numbers 11, 15, 2, 47, or any combination of them; doing things at exact times; model trains; helping people with their hobbies; or even Hildreth Milton Flitcraft. Rather, you see this *specific action* as obligatory only because you accept a more general disposition-maxim, "when God commands you to do something, you ought to do it." In this case, then, you accept the more specific Hildreth-helping maxim because you accept the more general God-obeying maxim.

3.1. Highest Disposition-Maxims and the *Gesinnung*

Kant does not say much about extremely specific disposition-maxims.[63] He does, however, say quite a bit about the broadest possible, or our "highest" (*Rel*, 6:25) disposition-maxims, which he reduces to two: the maxim of following the moral law and the maxim of following the law of self-love.[64] One can think of the disposition-maxim of

following the moral law as running, "in any circumstances, I ought to do my duty", or "in any circumstances, doing my duty is good" (call this "the Moral Maxim"), while the maxim of following the law of self-love might run, "in any circumstances, I ought to do what makes me happy" or "in any circumstances, doing what makes me happy is good"[65] (call this "the Prudential Maxim"); in Kant's words, "the principle of making {happiness} the supreme determining ground of choice is the principle of self-love" (*CPrR*, 5:22).

Kant thinks everyone accepts both these highest disposition-maxims: "The {moral} law … imposes itself on {the human being} irresistibly, because of his moral predisposition … He is, however, also dependent on the incentives of his sensuous nature because of his equally innocent natural predisposition, and he incorporates them too into his maxim (according to the subjective principle of self-love)" (*Rel*, 6:36). People generally judge the satisfaction of their sensible inclinations to be good, and also want to satisfy them, and also judge the satisfaction of their moral duties to be good, and want to satisfy them. In other words, it is the acceptance of these two highest disposition-maxims that determines what people see as valuable.

Although everyone accepts both highest disposition-maxims, a person's *Gesinnung* depends on which of these two highest disposition-maxims he makes supreme, i.e., the "condition" of the other: "the difference, whether the human being is good or evil, must not lie in the difference between the incentives that he incorporates into his maxim (not in the material of the maxim) but in their *subordination* (in the form of the maxim): *which of the two he makes the condition of the other*" (*Rel*, 6:36). Someone who makes obedience to the Moral Maxim conditional on following the

Prudential Maxim is evil, whereas someone who makes the Prudential Maxim conditional on the Moral Maxim is good.

It is not obvious what it means to subordinate one highest disposition-maxim to another. Take someone with a good *Gesinnung*, i.e., someone who makes promotion of her happiness conditional on the carrying-out of her moral obligations. Does she want and judge as good things in her self-interest except when they are immoral (at which point her judgment of their goodness and her desire to bring them about disappears)? Does she continue to judge actions that will make her happier as desirable and good, but simply refrain from undertaking them? Or does she do something else—say, continue to judge actions that are evil as good and desirable, only less good and less desirable than ones that are morally good?

We can rule out the first possibility. Kant is clear that someone with a good *Gesinnung* will still sometimes act immorally: "a human being, who incorporates … purity into his maxims, though on this account still not holy as such (for between maxim and deed there still is a wide gap), is nonetheless upon the road of endless progress toward holiness" (*Rel*, 6:46-47). So, the first interpretation of *Gesinnung*, according to which someone who is good cannot even commit an evil action (since she both judges it to be evil and has no desire to undertake it), is not right.

As I interpret him, Kant holds that anyone who willingly does something she knows to be immoral according to the moral law—i.e., someone who acts with "depravity" (*Rel*, 6:30)—is evil: "although with {depravity} there can still be legally good … actions, yet the mind's attitude is thereby corrupted at its root (so far as the moral *Gesinnung* is concerned), and hence the human being is designated as evil" (*Rel*,

6:30). However, a person who does immoral things out of weakness of will, or who acts from respect for morality, but "impurely", can still have a good *Gesinnung*: "This *innate* guilt {i.e., the evil *Gesinnung*} … can be judged in its first two stages (those of frailty and impurity) to be unintentional guilt" (*Rel*, 6:38).

Someone who exhibits "frailty" (*Rel*, 6:29)[66] (i.e., weakness of will) convinces herself that she is unable to resist her inclinations, and so succumbs to her immoral desire for happiness partially unwillingly. Someone who exhibits "impurity" has not "adopted the {moral} law *alone* as {her *Willkür*'s} *sufficient* incentive but … often (and perhaps always) needs still other incentives besides it in order to determine the *Willkür* for what duty requires" (*Rel*, 6:30). In other words, when a person acts impurely, she acts from respect for the moral law, but she is such that she also acts on sensible incentives. If the sensible incentives had not been there, she would not have acted in conformity with the moral law. It could also be that she acts impurely when she acts out of respect for the law against certain of her sensible inclinations, but is also such that if those sensible inclinations had been stronger, she would have abandoned acting out of respect for the law, because of frailty.[67]

One with a good *Gesinnung* judges immoral actions to be morally worse than morally right actions, but sometimes desires to do them more. It must be the case that she judges certain immoral actions to be at least somewhat good, otherwise she would have no motive to engage in them, and so would be unable to do so. For the same reason, she must also desire undertaking those immoral actions. However, she judges immoral actions to be less good than morally right actions because in order to engage in immorality she has to convince herself that she cannot resist her strong desires; that is, to

go ahead and freely engage in evil, she needs to first tell herself that she can do nothing else (she needs to excuse herself by telling herself that she is not free).[68] The fact that she has to invent this story, though, is evidence that she is, overall, aligned with the good; if she had an evil *Gesinnung* she would not need to come up with a tale of her lack of freedom. Instead, she would justify her conduct by adverting to something about her that entitles her to perpetrate actions that she would condemn others for.[69]

3.2. Nested Maxims

At least in the *Religion*, Kant thinks that there are highest maxims, and that these highest maxims somehow justify or influence the adoption of more specific disposition-maxims.[70] Unfortunately, he does not give reasons for thinking that there are highest maxims. In this section, I offer some considerations that might have convinced Kant to endorse the idea that there are highest maxims, and that these maxims influence the adoption of more specific maxims.

Kant's endorsement of highest maxims follows from his understanding of disposition-maxims in general. The reason people judge certain things as good and see them as desirable, Kant thinks, is that they accept disposition-maxims. So, if I see completing my dissertation as something I ought to do, and I desire to do it, this is because I accept the disposition-maxim, "when I have time, I ought to complete my dissertation". But it is not as though I judge writing my dissertation to be *primitively* good, for no explicable reason; perhaps this is how I judge the experience of pleasure, but not how I judge writing my dissertation. Instead, there are several reasons I can offer for wanting to complete my dissertation: because it is necessary for getting my Ph.D.,

because I have already spent so much time on it, because others will respect me for having done it, because it is fun, etc.

On Kant's view, there has to be an explanation for why each of these reasons count as reasons for me in the first place. Again, getting my Ph.D. is something that appeals to me, not because it is primitively good, but because it is a requirement for being able to get a tenure-track position in philosophy and because people will respect me for having it. And I care about each of these things for still deeper reasons; I care about getting a tenure-track position because it allows me to do things I enjoy doing, and I care about my peers' respecting me because their respect indicates to me that one of the things I most value has value for others.

Kant believes that the two most fundamental disposition-maxims are the Moral Maxim and the Prudential Maxim. These are the two disposition-maxims that determine everything else that a person takes to be of value. That is, things are valuable to a person only because they make her happy, or because they are morally required. Thus, all the other disposition-maxims a person accepts are shaped by these two highest disposition-maxims; if a person sees some action as good, it can only be because she thinks it will make her happy or because she thinks it is her duty.

I need to tidy up highest disposition-maxims by clarifying two things about them. First, what is the difference between the disposition-maxim, "in any circumstances, I ought to do what makes me happy" and "in any circumstances, I ought to do what gives me pleasure"? Second, I describe the Moral Maxim as, "in any circumstances, I ought to do my duty"; but should it not run, "in any circumstances, I ought to do my duty, *because it is my duty*"?

3.3. Happiness and Pleasure

Kant uses "happiness" in two ways. On the one hand, it is the idea of the rationally specified satisfaction of the largest possible set of inclinations,[71] and on the other hand, it can refer to the satisfaction of a person's occurrent desires:[72] "Happiness has two meanings: (1) the sum of all agreeable sensations. But this we cannot determine. (2) The gratification of all present inclinations" (*LM-M*, 29:899). By far Kant's more common practice is to use happiness in its first sense,[73] but it is no accident that he also specifies a second sense,[74] for by allowing happiness to have this second sense, he can claim that when are people are not acting from respect, they act always for the sake of happiness, even when they are doing something that they believe will make them less happy in the long run.

Even though happiness is an idea, it is an idea that everyone has and pursues: "all people have already, of themselves, the strongest and deepest inclination to happiness because it is just in this idea that all inclinations unite in one sum" (*G*, 4:399). Kant writes:

> the concept of happiness is such an indeterminate concept that, although every human being wishes to attain this, he can still never say determinately and consistently with himself what he really wishes and wills. The cause of this is that all the elements that belong to the concept of happiness are without exception empirical, that is, they must be borrowed from experience, and that nevertheless for the idea of happiness there is required an absolute whole, a maximum of well-being in my present condition and in every future condition. Now, it is impossible for the most insightful and at the same time most powerful but still finite being to frame for himself a determinate concept of what he really wills here. (*G*, 4:418)

One might reasonably wonder why Kant thinks, if happiness is an indeterminate idea, everyone formulates it and pursues it. Here Kant's hedonism comes into play. In the second *Critique* he writes:

> If the determination of {a human being's} will rests on the feeling of agreeableness or disagreeableness that he expects from some cause, it is all the same to him by what kind

> of representation he is affected. The only thing that concerns him, in order to decide upon a choice, is how intense, how long, how easily acquired, and how often repeated this agreeableness is. (*CPrR*, 5:23)

Everyone likes pleasure,[75] and every kind of action not motivated by respect is motivated by pleasure; moreover, every kind of sensible pleasure is commensurable. Consequently, whenever a person has to choose between satisfying two competing desires, but one promises more pleasure than the other, the person will judge the more pleasing one to be better – after all, she likes pleasure, and if she could have more of it without expending any effort, why not have more?

Call this principle of judgment ("choose the more pleasurable over the less pleasurable") the Pleasure Principle. The Pleasure Principle requires no special insight, but rather is obvious to any reasoning person, and is something to which everyone shows herself to be committed, either explicitly or implicitly, by the way she acts. And notice that the Pleasure Principle covers not just actions of little significance, like choosing between an apple and an orange, but also between choosing vacation spots, which college to attend, what career to have, whom to marry, and whether to have children. Indeed, if consistently applied it would lead to the most pleasurable life possible, that is, a life that contained the greatest set of satisfactions possible, which is Kant's usual understanding of happiness. In other words, someone who follows the Pleasure Principle is logically committed, regardless of whether she knows it, to wanting happiness most of all.

If anyone who prefers the more pleasurable over the less pleasurable thereby shows herself to be logically committed to wanting happiness most of all, then is the disposition-maxim, "in any circumstances, I ought to do what makes me happiest" the same as the disposition-maxim, "in any circumstances, I ought to do what gives me the most pleasure"? No, because the two disposition-maxims can come apart if someone does

what is more pleasurable in the short term but less pleasurable in the long term. Kant writes of this possible conflict:

> the precept of happiness is often so constituted that it greatly infringes upon some inclinations, and yet one can form no determinate and sure concept of the sum of the satisfaction of all inclinations under the name of happiness. Hence it is not to be wondered at that a single inclination, determinate both as to what it promises and as to the time within which it can be satisfied, can often outweigh a fluctuating idea, and that a man – for example, one suffering from gout – can choose to enjoy what he likes and put up with what he can since, according to his calculations, on this occasion at least he has not sacrificed the enjoyment of the present moment to the perhaps groundless expectation of a happiness that is supposed to lie in health. (*G*, 4:399)

Doing what you can to maximize your happiness can be quite difficult, because it can force you to give up a lesser, present satisfaction for a greater, future satisfaction. The problem is, since the greater satisfaction lies in the future, the chances of its coming to pass are less secure than the present satisfaction, which offers itself to you right now. As Kant puts it, "a man … can choose to enjoy what he likes … since … on this occasion at least he has not sacrificed the enjoyment of the present moment to the perhaps groundless expectation of a happiness that is supposed to lie in health." In other words, it is possible to choose the lesser pleasure over the greater, but the reason it is possible is that the greater one might not come to pass.

Thus, it appears that on Kant's view the reason it is possible to choose lesser over greater pleasure is that *one cannot be sure that the greater pleasure will in fact be greater*—it might not come to be at all. This seems to be rather different from choosing a lesser over a greater pleasure out of weakness of will. If this is the only reason why someone would choose a lesser over a greater pleasure, then "in any circumstances, I ought to do what gives me the most pleasure" ends up the same as "in any circumstances, I ought to do what gives me the most happiness"—after all, a putatively greater pleasure that will not come to pass will end up giving one less pleasure, and therefore will

contribute less to one's overall happiness, than a lesser pleasure that is assured; consequently, one who selects the lesser pleasure for this reason is really just choosing sensibly from a prudential point of view.

However, I think Kant's considered view is that people use the indeterminacy of the concept of happiness to deceive themselves into thinking that they are actually maximizing their happiness even though in clear moments they would concede their choice is a bad one. Notice that Kant writes, "one can form no determinate and sure concept of the sum of satisfaction of all inclinations under the name of happiness. Hence it is not to be wondered at that a single inclination … can often outweigh a fluctuating idea". His language suggests that he is explaining a phenomenon that must initially appear puzzling—namely, the apparent phenomenon of people choosing what they believe offers less pleasure over something that offers more. His explanation is that people can choose that which even they believe is less pleasant by temporarily deceiving themselves into thinking that it is in fact more pleasant. This is the sad consequence of guiding oneself by an idea that offers indeterminate guidance; even though happiness is what everyone most wants, the nature of the idea means that sometimes one fails at living up to one's strongest desire.[76]

The Prudential Maxim thus runs, "in any circumstances, I ought to do what makes me happy." And even though people can choose the course of action they believe will overall make them less happy, to do so they have to momentarily convince themselves that it will make a greater contribution to their happiness than its probably more pleasing alternative. Thus, I conclude that whenever people do not act on the Moral Maxim, they show their commitment to the Prudential Maxim.

3.4. The Moral Law and Reasons for Its Adoption

In this section I investigate whether one should describe the Moral Maxim as, "in any circumstances, I ought to do my duty" or "in any circumstances, I ought to do my duty, because it is my duty".

We can solve this issue by looking at the structure of maxims in general. Take a fairly specific disposition-maxim, such as "when I can help my well-connected friend, I ought to do so" or "when I can help my well-connected friend, doing so is good". This maxim of benevolence can be adopted for two reasons: either because it is morally required, or because it makes you happy. It is worth investigating, though, *why* helping your well-connected friend makes you happy. It could be because you are so constituted that helping people simply makes you happy; alternatively, it could be that you believe helping your well-connected friend will redound to your benefit in the future, so helping him seems good to you because it is a means to future happiness. If you adopt the maxim of benevolence out of happiness, then your adoption of the charity-maxim is partially explained by your commitment to the law of self-love. On the one hand, you must first of all be constituted such that helping your friend makes you happy; but if this benevolence is the kind of thing that *does* make you happy (and you know this), then you *will* adopt the charity-maxim (though you may or may not act on it).[77] After all, it makes you happy, so, given your allegiance to the law of self-love, you will judge it to be good; moreover, once you know it makes you happy, you will desire it—and the combination of your desire and your judgment of goodness amounts to the adoption of this disposition-maxim.

Note, though, that the maxim of benevolence runs, "when I can help my well-connected friend, I ought to do so", not "when I can help my well-connected friend, I

ought to do so, because it will redound to my benefit". Thus, when someone adopts one disposition-maxim because of a more fundamental disposition-maxim, the more fundamental disposition-maxim does not appear in the contents of the more superficial disposition-maxim.[78] This is not to say that the person cannot know that something may appear good to her because of the happiness she expects from it; as I will show in section 4, people can tell the difference between something's appearing good because they expect it to make them happier and because they think it is morally required. What I mean when I write that more fundamental disposition-maxims do not appear in the contents of less fundamental disposition-maxims is that when there is more than one available explanation for why a person thinks something will give her happiness, she will not necessarily know which of those explanations is the one that moves her, especially if she has a self-serving reason for thinking it is one explanation rather than the other.

Assuming this is right for typical disposition-maxims, there is reason for thinking it applies to highest disposition-maxims as well. Thus, just as the law of self-love does not show up in the charity-maxim, so the moral law does not show up in its own formulation. People endorse the moral law as their highest maxim not from duty, but rather because they are rationally committed to endorsing it.[79] In Kant's words, "it is not *because the law interests* us that it has validity for us … instead, the law interests because it is valid for us as human beings, since it arose from our will as intelligence and so from our proper self" (*G*, 4:460-61). I should also note that "because I am rationally committed to it" is not itself a reason that convinces anyone to endorse the moral law; rather, we can deduce the fact that people are committed to the moral law in virtue of the way they act and think about the world.

There is good textual evidence for thinking that Kant takes this line with regard to the moral law. In the *Religion* Kant writes, "Whenever we … say, 'The human being is by nature good,' or, 'He is by nature evil,' this only means that he holds within himself a first ground (to us inscrutable) for the adoption of good or evil (unlawful) maxims" (*Rel*, 6:21). One could argue that this passage means only that a person's decision to *subordinate* one highest maxim to the other is one for which we cannot provide a reason. While the passage does mean this, I do not think it means *only* this, because if the Moral Maxim ran, "in any circumstances, I ought to do my duty, because it is my duty" (or if the Prudential Maxim ran, "in any circumstances, I ought to make myself happy, because it will make me happy"), then we *would* have an explanation for why a person endorses one or the other highest maxim as her *Gesinnung*: either because it was her duty, or because it made her happy.

4. The Opacity of Maxims

So far we have established that there are two kinds of maxims, disposition- and policy-maxims; that every action is action on a disposition-maxim (the Maxim-Action Thesis); that every maxim consists of two elements, an incentive and a motive; that disposition-maxims can apply to extremely specific or extremely general situations; and that every maxim a person adopts is adopted either because of her commitment to the law of self-love or because of her commitment to the moral law. There is one last question to address: how much do people know about their own maxims?

There is some reason to think that Kant does not think we know very much about our own maxims at all. In the *Religion* he writes, "we cannot observe maxims, we cannot do so unproblematically even within ourselves" (*Rel*, 6:20). Similarly, in the *Groundwork*

he says, "it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty" (*G*, 4:407).

Neither statement implies that we do not know our own maxims. The passage from *Religion* says only that we cannot know others' maxims, not that we cannot know our own. Admittedly, it says we cannot know our own maxims "unproblematically", but this does not imply that we cannot know our own maxims *at all*. Similarly, in the *Groundwork* he claims that we cannot conceive of any situation in which we could justifiably be certain that someone performed an action in conformity with duty simply out of respect for the moral law. But from this we cannot conclude that we do not know when someone does something just out of the desire for happiness; nor does it force on us the view that we do not know when someone acts from respect for the law. All it says is that we cannot ever be certain that someone does something *just* from respect.

On my interpretation of disposition-maxims, we can know quite a bit about our own disposition-maxims. When we consciously deliberate before acting we can (often) know both what we desire (the incentive) and what we judge to be good (the motive). This is true even when we do not deliberate before acting; even when we act mindlessly, we often know after the fact what it was we desired, and also what we judged to be good (indeed, since a person cannot act without having a motive and a desire, we know that whenever someone acts to acquire something she desires that she also judges it to be good). This is true not just of those disposition-maxims that motivate our actions, but also of the disposition-maxims that we do not act on. Even if we do not act to realize

something we desire and judge good, we nonetheless often know what we desire and what we judge to be good.

Thus, we usually know the constituents of our disposition-maxims. But in many cases we know more than this; for instance, we know not just what we desire and judge good, but we also often know *why* we judge it good; at least, we know whether we judge it good because we believe it will promote our happiness or because we think it is morally required. Anyway, this must be Kant's view, for he claims that each person's conscience acts as a moral warning bell: if one considers doing something that is morally impermissible, one's conscience arises within her to tell her not to do so.[80] Conscience, in this capacity, is the "examining conscience" (*LE-V*, 27:615-17). The fact that the examining conscience exists, though, means that an agent can detect, at least some of the time, when she judges something to be good because it promotes her happiness and when she judges it morally good.

The conscience does not arise only before the fact, though; it also appears after someone acts, in which case it is the "judging conscience" (*LE-V*, 27:616), to pronounce her guilty or innocent of moral wrongdoing. Again, if a person's conscience can judge her after she has done something immoral, it follows that she at least has beliefs about the disposition-maxims of her past actions. Moreover, since Kant thinks people should abide by the judgments of their judging consciences,[81] he must think that the conscience is generally a reliable faculty[82] (except when it has been corrupted[83]).[84]

Not only can we know the elements of our disposition-maxims, we can also know when we are *considering* acting from self-love rather than respect (this is when the examining conscience warns us), and when we *actually* act from self-love instead of

respect (this is when our judging conscience condemns us). Moreover, we can know when we act from respect. One reason for thinking this is that respect is such a peculiar feeling compared to sensible incentives: "though respect is a feeling, it is not one *received* by means of influence; it is, instead, a feeling *self-wrought* by means of a rational concept and therefore specifically different from all feelings of the first kind, which can be reduced to inclination or fear" (*G*, 4:401n).[85] Indeed, it must be the case that we can know, not only when we immorally act out of self-love, but also when we *overcome* our inclinations for the sake of morality. If we did not know this, it would be impossible to feel self-contentment for acting virtuously: "When a thoughtful human being has overcome incentives to vice and is aware of having done his often bitter duty, he finds himself in a state that could well be called happiness, a state of contentment and peace of soul in which virtue is its own reward" (*MM*, 6:377). What you cannot know is whether you acted *only* from respect,[86] or whether *you would have* acted from respect had your temptations been greater.[87]

Because you cannot know whether you act from respect alone,[88] or whether you would have acted from respect if cooperating sensible incentives had been removed (or rebellious sensible incentives had been added), it follows that you cannot ever know whether your *Gesinning* is good.[89] Kant writes, "human beings cannot form themselves any concept of the degree and the strength of a force like that of a *Gesinnung* except by representing it surrounded by obstacles and yet – in the midst of the greatest possible temptations – victorious" (*Rel*, 6:61). Unfortunately, you cannot know whether you would be victorious.

We should remember that the exhibition of frailty is compatible with having a good *Gesinnung*. Consequently, when Kant says that we can never be sure that we have a good *Gesinnung* because of the counterfactual possibility that we might succumb to our inclinations, it cannot be that he thinks the bare fact of our possible frailty shows us to possibly have an evil *Gesinnung*. Rather, it must be that, for all we know, enough pleasure might lead us not only to convince ourselves that it is irresistible, but that we should join forces with it. That is, the promise of enough happiness can move a person to radically reconceive of her status in relation to the moral law.[90]

---

[1] I shall use "person", "persons", and "people" to refer to human beings (rather than beings such as angels or God), unless otherwise specified.

[2] See *Rel*, 6:22n and 25.

[3] See also *G*, 4:401n, *LE-M*, 29:603 and 608, *LE-V*, 27:495, and *MM*, 6:225 for definitions quite close to [M1]-[M3].

[4] Lewis White Beck notes, however, that a perhaps better translation for this term would be "spring":

> Abbott translates *Triebfeder* as 'motive' or 'spring.' 'Spring' follows a usage going back to the early seventeenth century, but not common now. There is good etymological justification for it, since *Feder* refers, e.g., to the mainspring of a watch. … the meaning of *Triebfeder* is obvious to a German, while *incentive* must be explained to a reader of English. It does not seem possible to find an entirely suitable English equivalent, and I suspect that the reason for this is that Kant himself did not use the word univocally. (Beck 1960, 91n)

I shall, however, follow customary usage and use "incentive".

[5] See Allison 1990, 5.

[6] For an argument that the notes taken on Kant's lectures are generally trustworthy, see Stark 2003, 16-19. Admittedly, he argues there only for the reliability of the student notes on Kant's lectures on *anthropology*, but considerations (4) and (5), which have to do, respectively, with the comparatively superior memorial abilities of 18th-century Prussians (especially when compared to contemporary Americans) and the high degree of corroboration from Kant's *Reflexionen* and his other lectures on metaphysics, as well as his published works, support the dependability of student manuscripts of Kant's lectures on metaphysics. Moreover, it should be noted that much of Kant's lectures on empirical psychology, which is where I get much of my information about his views on incentives, were later incorporated into his lectures on anthropology (see Louden 2000, 63 and 199n8).

[7] For examples of people who deny the Maxim-Action thesis, see Kitcher 2003, 236 and Frierson 2005, 16n44.

[8] See especially *MM*, 6:213, but also *CPrR*, 5:9n

[9] However, we must allow for behavior resulting from "affects". An affect is "the feeling of a pleasure or displeasure in the subject's present state <u>that does not let him rise to *reflection*</u> (the representation by means of reason as to whether he should give himself up to it or refuse it)" (*Ant*, 7:251). While the possibility of affects renders false Kant's bald assertion that a human *Willkür* is *never* determined to action without the mediation of reason, affects are atypical and fleeting:

> Affect is surprise through sensation … Affect is … rash, that is, it quickly grows to a degree of feeling that makes reflection impossible … What the affect of anger does not accomplish quickly it does not do at all … Affect works like water that breaks through a dam … Affect works on our health like an apoplectic fit. (*Ant*, 7:252).

[10] See this passage from the mid-1770s: "With all non-rational animals the stimuli have necessitating power, but with human beings the stimuli do not have necessitating power, but rather only impelling" (*LM-L*$_1$, 28:255); and also this one from 1794-95: "*human power of choice* … is never pure, *but rather always affected*. But the coeffecting stimuli can never determine it, but rather merely affect it sensibly, and in order to determine it there remains necessary the concurrence of the understanding" (*LM-K*$_3$, 29:1015).

[11] See also:

> An impelling cause contains actually the ground of the desire, an impelling cause is either a motive or a stimulus. Each impelling cause, when it is considered subjectively, is called an incentive of the soul, but when it is taken objectively, is called a stimulus. … Should the will be moved by stimuli or by motives? The stimuli spoil everything. (*LM-L*$_2$, 28:587)

According to this passage, "incentive" names an inclining mental state, and "stimulus" names the object that caused the incentive. Note that it is not Kant's usual practice to distinguish incentives and stimuli.

[12] Arguably, intellectual incentives can sometimes determine us, not just affect us, but that does not undercut the Incentive-Action Thesis.

[13] Kant says that people have to learn to act on maxims at *E*, §§72 and 78 as well.

[14] See also *E*, §§7, 79, and 93.

[15] He also writes in his lectures on anthropology that "people {without character} are like soft wax, every instant they take up another rule" (*LA*, 25:631/Frierson 2006, 628; translation Frierson's).

[16] See *MM*, 6:477.

[17] For more on this point, see Frierson 2006, 629-30.

[18] Both Rüdiger Bubner (Bubner 2001, 245-47) and Manfred Kuehn (Kuehn 2001, 144-49; see esp. 146) take policy-maxims to be the only kind of maxim that Kant admits.

[19] Some other passages further support the identification of dispositions and maxims in Kant: "We find ourselves enjoined by a moral law to good dispositions, as principles of our actions" (*LE-C*, 27:317); "fidelity in promises and benevolence from basic principles (not from instinct) have an inner worth … {that} does not consist in the effects arising from them … but in dispositions, that is, in maxims of the will that in this way are ready to manifest themselves through actions" (*G*, 4:435); and "The depravity of human nature is … not … a disposition (a subjective *principle* of maxims) to incorporate evil *qua evil* for incentive into one's maxim" (*Rel*, 6:37).

[20] See also *Rel*, 6:66:

> the human being's moral constitution ought to agree with this holiness. The latter must therefore be assumed in his *Gesinnung*, in the universal and pure maxim of the agreement of conduct with the law, as the germ from which all good is to be developed – [in a *Gesinnung*] which proceeds from a holy principle adopted by the human being in his supreme maxim.

[21] See also *LE-V*, 27:704: "decision in regard to the moral disposition of the agent is impossible. Here it is a matter, not of external actions, which are the object of judicial sentence {before an external or legal tribunal}, but of knowing the agent's motives".

[22] Kant also writes, "a human being who, from the time of his adoption of the principles of the good and throughout a sufficiently long life henceforth, has perceived the efficacy of these principles on what he does … has cause to infer, but only by way of conjecture, a fundamental improvement in his *Gesinnung*" (*Rel*, 6:68); and "our inference {about what *Gesinnung* we have} is drawn from perceptions that are only appearances of a good or bad *Gesinnung*" (*Rel*, 6:71).

[23] Kuehn denies that dispositions are identical to maxims, instead seeing them as "the motivation, or motivations expressed in our maxims" and as "the motivational aspect of maxims" (Kuehn 2001, 368). Given that Kuehn understands maxims exclusively as policy-maxims, I think this is the only interpretation of dispositions available to him. While there is much to be said for this interpretation of dispositions (regardless of whether one believes that Kant uses "maxim" univocally or equivocally), it seems to me that it would reduce dispositions to incentives.

[24] What I call "disposition-maxims" is very close to how Talbot Brewer understands maxims in Brewer 2002. Brewer, however, is noncommittal about whether Kant actually understands maxims in this way; he merely thinks Kant should have understood maxims in this way ("I am more interested in whether the view I've sketched is plausible than in whether it is properly called Kantianism" (Brewer 2002, 568)).

[25] Incentives can also make a person want to avoid a state of affairs, but for simplicity's sake I take incentives only to be psychological states that make a person want to bring about a state of affairs. See also endnote 4.

[26] See endnote 4.
[27] See, e.g., *LM-L*$_1$, 28:254-55, *LM-M*, 29:895, *LE-C*, 27:267, *G*, 4:427-28, *CPrR*, 5:72, *LM-L*$_2$, 28:587-88, and *LE-V*, 27:493-94.
[28] However, this passage, from the mid-1770s, comes close: "Understanding submits motives for omitting some action; sensibility, on the contrary, stimuli for committing it. But this dispute ends either when the stimuli no longer drive [us] (then the higher faculty triumphs, and the motives are predominant); or if the understanding submits no motives at all, then sensibility becomes predominant" (*LM-L*$_1$, 28:256). See also endnote 51.
[29] Other passages that define incentive and motive are: *LM-L*$_1$, 28:254-55 and 256, *LE-C*, 27:256-57, 268, and 1429, *G*, 4:427, *CPrR*, 5:72, *LE-V*, 27:493, and *LM-K*$_3$, 29:1015 and 1016.
[30] Here I follow Frierson 2005, 9. See also *LM-K*$_3$, 29:1012-13.
[31] See also *Rel*, 6:29, *MM*, 6:212 and *Ant*, 7:265.
[32] See esp. *G*, 4:425; *C2*, 5:72-73, 75, and 79; and *MM*, 6:379, 380n, 380-81, and 383.
[33] For example, "the force in you that strives only toward happiness is *inclination*" (*MM*, 6:481); "Sexual inclination is also called '*love*' … and is, in fact, the strongest possible sensible pleasure in an object" (*MM*, 6:426)
[34] I thus reject the argument of Reath 2006B that Kant is not a hedonist but a preference-satisfaction theorist (note that Reath himself concedes that the 1989 version of that argument—which relied on the dubious interpretative move that Kant's talk of pleasure preceding desire has only to do with how we came to acquire the desires in the first place—is flawed; see Reath 2006B, 38, 56-59 and 61n9 and also Herman 2007). I discuss the plausibility of Kant's hedonism in chapter 5, sections 3.3.2.2 and 5.
[35] See also *LM-L*$_1$, 28:254 and *LM-M*, 29:878 and 900.
[36] See also *LM-L*$_1$, 28:229 and 254-55, *LM-L*$_2$, 28:586, and *MM*, 6:212-13.
[37] See also the sources mentioned in endnote 17 and *LE-C*, 27:267 and *LM-D*, 28:677, quoted in section 2.1.
[38] Kant also writes, "The power of free choice is determined by motives. Since these originate only in the understanding, they are intellectual impelling causes. These are the concepts of the good" (*LM-M*, 29:896). Context, however, indicates that "motives" refers here particularly to moral motives rather than to motives in general.
[39] See also *LE-C*, 27:257.
[40] See also *LE-C*, 27:257 and 268, and *G*, 4:389.
[41] Thanks to Timothy Rosenkoetter for pointing this potential source of confusion, as well as for suggesting the locution, "making practically relevant".
[42] See also *LE-C*, 27:255-56, *G*, 4:414, and *LM-L*$_1$, 28:257-58.
[43] I accept the "Consequent Scope" interpretation of hypothetical imperatives offered by Schroeder 2005 (instead of the "Wide Scope" interpretation). Schroeder writes:

> According to Consequent Scope, hypothetical imperatives enjoin *those with the end* to take the necessary means. Their mandate is, as Kant puts it, 'based on a presupposition', and they don't mandate it to those who don't satisfy the presupposition. But according to Wide Scope, the Hypothetical Imperative is not based on any presupposition. It applies to everyone unconditionally, no matter what they are like. It sounds, in short, very much like a special kind of *categorical* imperative. (Schroeder 2005, 358)

The main worry about interpreting Kant as a Consequent Scope theorist is that the Consequent Scope theory allows for things to count as obligations that clearly should not be obligations: "Take the case of someone who wills to be an axe-murderer. If we accept Consequent Scope, then all we have to do is to apply *modus ponens* to yield the conclusions that he ought to sharpen his axe and that he ought to stake out victims" (Schroeder 2005, 368). As Schroeder goes on to point out, Kant avoids this consequence because he does not think that immoral ends can actually be willed. This is not to say that he thinks people cannot do immoral things, but only that they cannot will them with their *Wille*: "If an end doesn't pass the test of the Categorical Imperative, then it can't be *willed.* … the Categorical Imperative tells us … what things could possibly be products of our *wille* {*sic*} … though bad ends can be set by your *willkür* {*sic*} and thus be the product of your *choice*, they can't be the product of your *wille* {*sic*}" (Schroeder 2005, 369).
[44] With the Categorical Imperative "we have arrived, within the moral cognition of common human reason, at its principle, which it admittedly does not think so abstractly in a universal form but which it actually has always before its eyes and uses as the norm for its appraisals" (*G*, 4:403-4).

[45] "{W}e cannot be determined by mere representations of reason; it must also give us incentives, and these it also gives us, because our conscience approves or disapproves" (*LM-M*, 29:900).
[46] In sections 4 and 6 of chapter 5 I show how an evil person can be blind to the deliverances of the moral law within her.
[47] See *LE-C*, 27:255-56.
[48] See especially *CPrR*, 5:58-59n, where Kant distinguishes between two etiologies of the perception of goodness: "we represent to ourselves something as good when and *because we desire* (will) *it*, or also: we desire something *because we represent it to ourselves as good*".
[49] My analysis of respect is rather close to that found in Reath 2006A.
[50] Hence Kant's remark that "sensible feeling, which underlies all our inclinations, is indeed the condition of that feeling we call respect, but the cause determining it lies in pure practical reason" (*CPrR*, 5:75).
[51] We are now positioned to explain Kant's remark that "Understanding submits motives for omitting some action; sensibility, on the contrary, stimuli for committing it" (*LM-L*$_1$, 28:256; see also endnote 28). The understanding always submits motives for omitting some action, even in cases where someone performs a positive duty such as self-improvement or benevolence to others out of respect, because in order for someone to do anything from respect, she has to resist either a countervailing inclination, or resist performing the action for the wrong reason.
[52] See also *CPrR*, 5:75.
[53] To be fair, Kant, at *CPrR*, 5:77-78, says that respect involves neither pleasure nor displeasure; by this he simply means that it does not involve *sensible* pleasure or displeasure. But it clearly does involve a kind of *intellectual* pleasure and displeasure (e.g., "we can see a priori that the moral law, as the determining ground of the will, must by thwarting all our inclinations produce a feeling that can be called pain" (*CPrR*, 5:73)).
[54] Kant writes, of both sensible and intellectual pleasure, that "Pleasure is when a representation contains a ground for becoming determined, for producing again the same representation, or for continuing it when it is there" (*LM-L*$_2$, 28:586).
[55] One might wonder how important motives really are if Kant discusses them mostly in his lecture notes and only a couple of times in his central ethical works. I think, though, that he discusses motives quite a bit in his central ethical works, under the guise of "interests." I take interests to relate to motives as inclinations relate to incentives (see section 2.2.3.1): just as an having an inclination to φ means either that you regularly φ or that the opportunity to φ presents you with incentives, so having an interest in φ-ing means that you judge φ-ing to be good, even when you do not right now want to φ. Having an interest in exercising means not only judging exercising to be good just before I engage in it, but also being prepared to make room for exercising in my life.

Here are some textual considerations in favor such an interpretation. First, in the *Groundwork* Kant writes, "An interest is that by which reason becomes practical, i.e. becomes a cause determining the will. Hence only of a rational being does one say that he takes an interest in something; nonrational creatures feel only sensible impulses" (*G*, 4:460n); if interests relate to motives in the way I just sketched, then this passage makes sense: animals do not take interests in things because they do not have motives but are instead determined to action only through incentives. He also writes:

> Reason takes an immediate interest in an action only when the universal validity of the maxim of the action is a sufficient determining ground of the will. Only such an interest is pure. But if it can determine the will only by means of another object of desire or on the presupposition of a special feeling of the subject, then reason takes only a mediate interest in the action, and since reason all by itself, without experience, can discover neither objects of the will nor a special feeling lying at its basis, this latter interest would be only empirical and not a pure rational interest. (*G*, 4:460n)

Again, this mirrors the story I have told about moral and pragmatic motives. In the case of immediate interests (moral motives), reason determines the will directly (i.e., itself produces an incentive), whereas in the case of mediate interests, reason determines the will only after a person has a desire for some object.

Second, in the *Critique of Practical Reason* he writes:

> From the concept of an incentive arises that of an *interest*, which can never be attributed to any being unless it has reason and which signifies an *incentive* of the will insofar as it is *represented by reason*. Since in a morally good will the law itself must be the incentive, the *moral interest* is a pure sense-free interest of practical reason alone. On the concept of an interest is based that of a

maxim. A maxim is therefore morally genuine only if it rests solely on the interest one takes in compliance with the law. (*CPrR*, 5:79)

Sensible interests signify sensible incentives "represented by reason"—that is, a sensible interest, which depends on pragmatic motives, represents a state of affairs for which someone has a desire as good. A moral interest, though, is "a pure sense-free interest of practical reason alone"—that is, it is not a representation of an incentive, but gives rise to an incentive. Again, this mirrors the story I told about motives.

For all I have said so far, though, interests could simply *be* motives. This passage from the second *Critique*, though, suggests the relationship I have drawn above: "To every faculty of the mind one can attribute an *interest*, that is, a principle that contains the condition under which alone its exercise is promoted. Reason, as the faculty of principles, determines the interest of all the powers of the mind but itself determines its own" (*CPrR*, 5:119-20). If interests were motives, it would be strange to say that faculties of reason had motives, i.e., endorsed particular imperatives. If having an interest, though, amounted to be being disposed to judge certain things as good when they became salient, then this passage makes more sense.

Finally, there are a number of passages on interests in the *Critique of the Power of Judgment*. See *CJ*, 5:204-10.

[56] My account of disposition-maxims is somewhat close to that of Patricia Kitcher 2003 and Richard McCarty 2006.

[57] See also *LE-C*, 27:268, 274-75 and *LE-V*, 28:493-94 (though it should be noted that at *LE-C*, 27:274-75 Kant seems to treat respect itself as a motive rather than the result of accepting a motive).

[58] Notice that Kant writes that a maxim "is the subjectively practical principle, insofar as the subject makes the rule by which he is to act into the motive of his action as well"; this supports my interpretation of motives as endorsements of imperatives.

[59] Alternatively, "in C, bringing about state of affairs S is good" or "in C, I ought to bring about S".

[60] Kant clearly thinks that we can have incentives without having motives; he also thinks we can judge something to be good without desiring it, as is the case with aesthetic judgments (see, e.g., *CJ*, 5:205n).

[61] See *MM*, 6:431.

[62] I take this name, and the idea of an extremely specific maxim, from Wood 1999, 102-05.

[63] In the *Groundwork*, though, he examines this maxim: "from self-love I make it my principle to shorten my life when its longer duration threatens more trouble than it promises agreeableness" (*G*, 4:422). Although this maxim can apply to a wide range of situations, in most agents' lives they will be able to act on it only once; thus, it is pretty specific for a maxim.

[64] See *LE-C*, 27:422 and *Rel*, 6:36.

[65] Cf. Korsgaard 1996, 165, and Kitcher 2003, 226.

[66] See also *LE-C*, 27:293-95, where he discusses "weakness" (which he renamed, in 1793, "frailty") and "frailty" (which he renamed, in 1793, "depravity"), and also *LE-V*, 27:570.

[67] Note that if she acts out of respect for the moral law, but would have acted immorally if competing sensible incentives had been stronger, and moreover *would have acted immorally not out of weakness, but willingly*, then she shows herself to have an evil *Gesinnung*, even as she acts out of respect for the moral law.

[68] See *CPrR*, 5:99.

[69] I discuss these strategies in greater detail in chapter 5, section 6.1.

[70] See *Rel*, 6:21, 21n, and 25.

[71] "[H]appiness is the maximal degree of satisfaction of all our inclinations" (*LE-M*, 29:598).

[72] In *The Metaphysics of Morals* Kant defines an "*end* is an object of the choice (of a rational being), through the representation of which choice is determined to an action to bring this object about" (*MM*, 6:381), but in his lectures on *The Metaphysics of Morals* claims that ""end" is equivalent in meaning to the concept of what brings happiness" (*LE-V*, 27:544). Arguably, he is simply saying that every object of choice is meant to contribute to happiness, but I think it is just as likely that achieving any end, at least any material end, brings happiness.

[73] He uses it in this sense at *LE-C*, 27:366-67, *LM-V*, 28:446, *LE-M*, 29:598, *G*, 4:399, 405, and 418, *CPrR*, 5:22, 25, 61, and 73, *CJ*, 5:208 and 434n, *LM-L*$_2$, 28:593, *OCS*, 8:282, 283, and 290, *LE-V*, 27:499 and *Rel*, 6:58 and 67.

[74] He uses it in this sense at *LE-M*, 29:623, *CPrR*, 5:23 and 25, *LE-V*, 27:544, and *MM*, 6:387.

[75] Everyone likes pleasure because pleasure is evidence that one's life is being promoted: "The feeling of the promotion of life is pleasure" (*LM-L*$_2$, 28:586). See also *LM-L*$_1$, 28:247, *LM-M*, 29:890-91 and 894, *CPrR*, 5:9n, and *Ant*, 7:231. I talk about the relationship between pleasure and life in greater detail in chapter 5, section 3.3.2.2.

[76] As Kant puts it, "it is a misfortune that the concept of happiness is such an indeterminate concept that, although every human being wishes to attain this, he can never say determinately and consistently with himself what he really wishes and wills" (*G*, 4:418).

[77] I do not mean to say that this adoption is caused; it is a free adoption, in the same sense that drawing the conclusion that 2+2 = 4 is a freely drawn conclusion.

[78] Hence Kant's statement that:

> one can never say that stimuli should not have affected it, e.g., with the giving of alms stimuli would already be there if he gives the alms away for the sake of the comfort that he draws from it, or from love of honor not to be harshly reprimanded in the eyes of his neighbors, or from compassion toward the tattered needy one, or toward his pleas touching the weakness of the giver, or also for an expected reward from God: the stimuli are often so hidden that one must examine oneself closely (*LM-K*$_3$, 29:1015).

[79] Here I follow Johnson 2007, 13-14.

[80] "A distinction can be made between conscience before and after the act" (*LE-C*, 27:356).

[81] See *LE-C*, 27:353 and 356.

[82] See, e.g., *CPrR*, 5:35.

[83] I discuss what a corrupted conscience is like, and how it is corrupted, in sections 5 and 6.2 of chapter 5.

[84] Note also that Kant is quite explicit that conscience judges our disposition-maxims: "This *forum internum* {i.e., conscience} is a *forum divinum*, in that it judges us by our very dispositions, and we cannot, indeed, form any concept of the *forum divinum* other than that we must pass sentence on ourselves according to our dispositions" (*LE-C*, 27:297).

[85] See also *CPrR*, 5:33, 75-76, 80-81, 87-88, and 116-17.

[86] See *G*, 4:407.

[87] See *Rel*, 6:38 and *MM*, 6:392-93.

[88] See, in addition to *G*, 4:407, *LM-M*, 29:897: "We are conscious only of the incentives or stimuli which are clear representations. But we can also have obscure representations and stimuli for something of which we thus are not conscious." The possibility of unconscious incentives makes it impossible to ever be *certain* that anyone has ever acted *only* from respect.

[89] You can, of course, know when your *Gesinning* is evil; this is when you intentionally do something you know to be wrong. Unfortunately, people who intentionally do things they know to be wrong almost always have defective consciences that prevent them from seeing their full responsibility for their actions, at least at the time of their actions. As Kant says, an evil *Gesinnung* "is characterized by a certain *perfidy* on the part of the human heart in deceiving itself as regards its own good or evil disposition" (*Rel*, 6:38).

[90] I discuss this process in chapter 5, section 6.2.

Chapter 3: The Propensity to Evil and the *Gesinnung*

1. Introduction

In this chapter I explain Kant's concept of the propensity to evil. On my view, Kant uses "propensity to evil" equivocally, sometimes meaning it to refer to the mere propensity to evil itself, which mere propensity I call the "susceptibility to evil", and sometimes using it to refer to the inclination to evil to which the mere propensity gives rise, which evil inclination I call the "tendency to evil". The tendency to evil is a disposition (in the non-Kantian sense) to generate disproportionately strong, judgment-resistant desires. A bare propensity to evil or, what is the same thing, an idle propensity to evil, is a surefire disposition to develop a tendency to evil when one matures in an evil society; thus, it is a disposition to form another disposition, or a second-order disposition.

Although Kant often writes as though the propensity to evil *qua* tendency and an evil *Gesinnung* are the same thing, this is not what he thinks (though "tendency to evil" is the appellation I most commonly employ, I interchange it with "propensity to evil *qua* tendency" and "drive to evil"). Rather, an evil *Gesinnung* is the supreme maxim of subordinating the Moral Maxim ("in any circumstances, I ought to do my duty") to the Prudential Maxim ("in any circumstances, I ought to do what makes me happy").[1] Unlike the susceptibility to evil, which is congenital, evil *Gesinnungen* are freely adopted. However, it is only by adopting an evil *Gesinnung* that a person can "activate" her susceptibility to evil, causing it to produce the tendency to evil. It should also be noted that though an evil *Gesinnung* can be replaced by a good *Gesinnung*, a tendency to evil,

once generated, is permanent. Thus, anyone who has an evil *Gesinnung* has a tendency to evil, while anyone who has a tendency to evil has *or had* an evil *Gesinnung*.

Kant believes everyone has a tendency to evil (I call this claim the "Universality Thesis" (UT)); thus, even though he believes everyone can overcome her evil *Gesinnung*, he also thinks everyone begins her moral life by putting on an evil *Gesinnung*. My goal in the next chapter is to show what his reasons are for believing the UT. In this chapter, I restrict myself just to discussing the nature of the propensity to evil (both *qua* susceptibility and *qua* tendency) and its relation to one's *Gesinnung*. I focus on the tendency to evil because, unless one knows precisely how Kant understands it, it is impossible to piece together his argument for its ubiquity.

2. Propensities and Predispositions

In this section I argue that propensities, among them the propensity to evil *qua* susceptibility, are the same thing as what Kant in the *Religion within the Boundaries of Mere Reason* calls "contingent predispositions" (*Rel*, 6:28) (I use the noun phrases "susceptibility to evil", "propensity to evil *qua susceptibility*" and "idle propensity to evil" interchangeably). It is important to establish this, not only to justify my division of the propensity to evil into a susceptibility to evil and a tendency to evil, but also because the susceptibility to evil's contingency means that Kant does not need to offer a synthetic, *a priori* argument to establish the tendency to evil. After all, the only faculties or claims that require synthetic, *a priori* arguments to justify their existence or truth are ones that have the marks of necessity and universality.[2] Given that there are people who do not have a tendency to evil—viz., people who achieve rationality in a world exemplifying the highest good—, the propensity to evil *qua* tendency is neither universal nor necessary, so

Kant does not need to offer a synthetic, *a priori* argument for the claim that everyone has it (i.e. the UT).

Propensities

In *Anthropology from a Pragmatic Point of View*, Kant defines a propensity as the "subjective *possibility* of the emergence of a certain desire, which *precedes* the representation of its object" (*Ant*, 7:265). A propensity is what allows a person to develop inclinations[3] of a certain kind.[4] Someone with a propensity for *x* will develop an inclination (a habitual desire) for *x* if she is exposed to *x* in the right way (in contemporary parlance, a propensity is a probabilistic disposition to develop another disposition, or a second-order, probabilistic disposition).[5] By the same token, if one has a propensity for *x* but is never exposed to *x* in the right way, then one will still have a propensity for *x*, but it will be idle—that is, one will have no inclination for *x*. For instance, for a person to develop an inclination for alcohol, she has to have a propensity for alcohol; and for her to activate this propensity, she has to consume alcohol: "all savages have a propensity for intoxicants; for although many of them have no acquaintance at all with intoxication, and hence absolutely no desire for the things that produce it, let them try these things but once, and there is aroused in them an almost inextinguishable desire for them" (*Rel*, 6:29n).[6] If she does not consume alcohol, though, then she will not develop an inclination for it, even though she is still susceptible to generating one.

I need to note one thing about propensities right away. Kant's definition of propensity in the *Anthropology*,[7] the *Religion*,[8] and in his lectures on anthropology[9] suggests that the kinds of inclinations a person can have are limited by the propensities

she has. According to this interpretation, if someone does not have a propensity to alcohol, she should not be able to develop an inclination (i.e., a habitual K-desire[10] for) alcohol at all. But this is not right; while it may be true that "savages" usually develop an inclination for alcohol as soon as they try it, it is quite implausible to read Kant as saying that Europeans, who, in contrast to savages, do not have a propensity for alcohol, cannot develop an inclination for alcohol.

On the other hand, the example Kant presents of a propensity in the *Religion* suggests that propensities need not underlie all inclinations; just those inclinations we can develop after little exposure to those things for which they are inclinations. After all, why focus on savages and their alleged propensity for intoxicants unless this propensity distinguishes them from Europeans?

So, propensities appear to be necessary for developing *any* inclinations at all; yet we posit propensities only to explain inclinations that develop particularly suddenly and vivaciously. To explain how both these things can be true of propensities, it is useful to bring in predispositions.

Predispositions

Predispositions, like propensities, allow for the emergence of inclinations. Three predispositions—which Kant calls "original" predispositions (*Rel*, 6:28) or, collectively, "the original predisposition to good" (*Rel*, 6:26)—that Kant pays special attention to are the predispositions to animality, humanity, and personality.

The predisposition to animality gives rise to three general inclinations: the inclinations "for self-preservation", "for the propagation of the species, through the sexual drive, and for the preservation of the offspring thereby begotten through

breeding", and "for community with other human beings, i.e. the social drive" (*Rel*, 6:26). Kant does not say this, but we can assume that the inclination for self-preservation gives rise to desires not just to preserve one's life in the face of danger, but also for food, drink, and sleep (after all, food, drink, and sleep are essential to our continued existence as living beings).

All the inclinations that develop out of the predisposition[11] to humanity "can be brought under the general title of a self-love which is physical and yet *involves comparison* … Out of this self-love originates the inclination *to gain worth in the opinion of others*" (*Rel*, 6:27). In other words, the predisposition to humanity gives rise to several types of inclination, but what they have in common is that each produces desires for having one's lot in life compare favorably to others.

Kant writes that "The predisposition to personality is the susceptibility to respect for the moral law *as of itself a sufficient incentive to the power of choice*" (*Rel*, 6:27). Without the predisposition to personality, we would be incapable of being able to act out of respect for the moral law. If one were to describe the predisposition to personality as generating an inclination, that inclination would be the "moral drive" (or as Kant calls it, "moral feeling" (*MM*, 6:399)).

Like propensities, each of these predispositions gives rise to inclinations. However, they generate a wide variety of quite general inclinations, each of which has more specific inclinations under it. For instance, the predisposition to animality produces the inclination for self-preservation; under this inclination stands the inclination for food, and under this inclination stands the inclination for sweet things.

Noting this relationship among inclinations of varying levels of generality allows us to make sense of how it can be true both that a propensity is merely a disposition to produce an inclination given the right triggering conditions, and also how it is a disposition to have an inclination suddenly emerge with great vivacity (again, given the proper triggering condition). While it is true that everyone, European and savage alike, can develop an inclination for intoxicants, savages are genetically primed (pardon the anachronism) to develop an inclination for intoxicants in particular. That is, not only do savages, like typical Europeans, have a general predisposition for animality that mediately gives rise to an inclination for intoxicants (given enough exposure to intoxicants), they also have a *more specific* propensity for intoxicants that makes them susceptible to quickly developing a powerful inclination for intoxicants. So, while everyone (except those allergic to alcohol) can develop an inclination for alcohol thanks to their predisposition to animality, those who have a propensity for alcohol over and above their predisposition to animality can acquire an inclination for it that is not only stronger, but also easier to get. (This distinction between specific and general propensities will be important for understanding the propensity to evil *qua* tendency.)

Throughout this discussion of original predispositions, I have assumed a kinship between original predispositions and propensities. This is because both propensities and original predispositions allow for the emergence of certain kinds of inclinations, though the things Kant calls propensities (sometimes) generate more specific inclinations than original predispositions do. However, there are some dissimilarities between propensities and original predispositions. For one thing, original predispositions are second-order dispositions that everyone has, indeed, *must* have: the three predispositions "are *original*,

for they belong to the possibility of human nature. … By the predispositions of a being we understand the constituent parts required for it as well as the forms of their combination that make for such a being. They are *original* if they belong with necessity to the possibility of this being" (*Rel*, 6:28). Thus, a being who did not have either the predisposition to animality, humanity, or personality would not be a human person.[12]

Kant claims that original predispositions are necessary to people, and it seems likely that this necessity is conceptual. One may doubt, though, that when Kant talks of the necessity of original predispositions, he means that we must conceive of persons as having all three predispositions, on pain of not conceiving a person at all. This is because Kant cautions in a footnote that:

> We cannot consider {the predisposition to personality} as already included in the concept of the {predisposition to humanity}, but must necessarily treat it as a special predisposition. For from the fact that a being has reason does not at all follow that, simply by virtue of representing its maxims as suited to universal legislation, this reason contains a faculty of determining the power of choice unconditionally, and hence to be 'practical' on its own … Were this {moral} law not given to us from within, no amount of subtle reasoning on our part would produce it or win our power of choice over to it. (*Rel*, 6:26n)

Someone with a predisposition to humanity but no predisposition to personality could never act from respect for the moral law. One might take this to show that the predisposition to personality is not a part of the concept of the human being; after all, empiricists in the Humean tradition think that the character of morality can be explained without recourse to some special ability to act from respect. Kant, though, would simply disagree. He would hold that commonsense beliefs about morality commit us to positing both the moral law and the ability to act from respect for the moral law. Although Kant's conceptual analysis is debatable, it does not follow from this that he did not see the three original predispositions as part of the concept of persons.

Original predispositions and propensities are both predispositions, because both generate inclinations. But whereas original predispositions are necessary to the concept of a human person, propensities are not; instead, they are "*contingent*" predispositions (*Rel*, 6:28).

Although Kant clearly distinguishes between original and contingent predispositions, he never identifies contingent predispositions with propensities. Still, the language he uses in his definitions of propensity, at least in the *Religion*, suggests such an identity: "By *propensity* … I understand the subjective ground of the possibility of an inclination … insofar as this possibility is contingent for humanity in general" (*Rel*, 6:29); and "*Propensity* is actually only the *predisposition* to desire an enjoyment which, when the subject has experienced it, arouses *inclination* to it" (*Rel*, 6:29n).

Another reason aside from Kant's language for thinking that contingent predispositions and propensities are identical comes from asking what the difference would be between contingent predispositions and propensities, assuming they were not the same. Kant writes that a propensity "is distinguished from a predisposition in that a propensity can indeed be innate yet *may* be represented as not being such; it can rather be thought of (if it is good) as *acquired*, or (if evil) as *brought* by the human being *upon* himself" (*Rel*, 6:29), which suggests that people can be represented as being causally or morally responsible for their propensities but not for their predispositions, contingent or original.

But what does Kant mean when he says that we can represent propensities as acquired or brought upon oneself? I think Kant is saying that they are activated. That is, before a propensity for *x* will generate an inclination for *x*, an agent has to experience *x* in

some way. For instance, if you know you are genetically predisposed to alcoholism, but choose to drink alcohol and become an alcoholic as a result, there is a sense in which you bring alcoholism upon yourself: you bring alcoholism upon yourself by freely feeding your propensity for it, thereby allowing the inclination to flourish. Because every propensity requires an activating condition, anyone who has an inclination stemming from a propensity can be seen as responsible for it, at least in some sense.

The original predispositions, though, also require activation. Kant indicates in *Education* that while people start out with a full roster of (original) predispositions,[13] those predispositions do not begin fully developed, but instead have to be activated. Thus, he says, "There are many germs lying undeveloped in man. It is for us {educators} to make these germs grow, by *developing his natural gifts* in their due proportion, and to see that he fulfils his destiny" (*E*, §10).[14]

If the original predispositions require activation, there is no reason to think that contingent predispositions do not also require activation. But if contingent predispositions require activation, then it is hard to see any difference between them and propensities.[15] Thus, I conclude that they are the same.

The fact that original predispositions are activated, though, brings up another question: why cannot people be seen as responsible for their original predispositions? The answer is that original predispositions are always activated in the normal course of things. By "the normal course of things", I mean the following: in any human society in which a normal human being (i.e. someone who does not have some disorder, congenital or acquired, that prevents her from developing in the normal way) can mature, she will eventually activate all her original predispositions; if she does not get the chance to

mature, then it is still true that she would have enlivened all her original predispositions had she had the chance to mature. So, a child who unfortunately dies before becoming rational will never have the chance to activate her predispositions to humanity and personality (though she will have activated her predisposition to animality[16]), but it nonetheless remains the case that if she had not died, she would have developed those predispositions. In other words, because human beings *necessarily* activate their original predispositions, they should not be represented as responsible for them.

Let me summarize what I have shown so far. Both propensities and original predispositions are second-order dispositions, that is, dispositions that, given the right triggering conditions, produce dispositions to engage in certain kinds of behavior (i.e., inclinations). Original predispositions, though, give rise to very broad inclinations – inclinations for self-preservation, inclinations for sociability, and the inclination to act in conformity with morality out of respect for the law. Similarly, propensities are dispositions that, with the right activating conditions, can produce general inclinations; alternatively, they can be dispositions that suddenly generate powerful, usually more specific, inclinations. In either case, though, they are *contingent* predispositions – dispositions we could conceive of a person not having – while original predispositions are *necessary* predispositions – dispositions we could not conceive of a person without.

Finally, because propensities are contingent—a person may fail altogether to have a particular propensity, or she may refrain from activating it, even if she does have it—, Kant thinks we can represent people as responsible for them (more precisely, Kant thinks we can represent people as responsible for their *activation*). On the other hand, because original predispositions are necessary—because everyone has them, and they are always

activated in the normal course of things—we cannot represent people as responsible for their triggering.

It is time now to use the definitions we have developed to evaluate the propensity to evil.

3. The Propensity to Evil and *Gesinnungen*

Above, I wrote that someone with a propensity for *x* will develop an inclination for *x* if exposed to the right triggering conditions but will not develop an inclination for *x* if she never experiences it (though she will still have a propensity for *x*). By simple substitution, someone with a propensity to evil will develop an inclination for evil if exposed to the right triggering conditions, but will not acquire an inclination for evil if she is never exposed to evil in the right way. Because someone can have a propensity to evil without having an inclination for evil, I characterize the propensity to evil in two ways. First, the bare propensity to evil itself is a susceptibility to evil, because even if it is idle, a person still has it, and still could develop an inclination to evil if put in the right circumstances. Second, the inclination to evil to which the propensity to evil gives rise is the tendency to evil. I call the inclination to evil a tendency to evil because it is not so much a desire for a particular kind of thing as it is something that modifies all of a person's sensible inclinations.

Because Kant never explicitly distinguishes between the mere susceptibility to evil and the tendency to evil, there is no explicit definition of either available. This is unfortunate, because Kant is committed to what I above call the Universality Thesis (the UT). According to the UT, everyone has a drive to evil (given what we have so far said about the tendency to evil, the UT amounts to the claim that everyone's susceptibility to

evil is activated, thus producing in each the tendency to evil); if one is not careful, though, one might take the UT to be simply the claim that everyone has a propensity to evil. Elucidating Kant's argumentation for the UT is the task of chapter 4, but in order to grasp that reasoning, one has to understand precisely what Kant means to establish, i.e. the tendency to evil.

On my interpretation, the tendency to evil is a property the possession of which strengthens our sensible desires such that they become disproportionately strong and judgment-resistant. By describing the sensible desires that the drive to evil touches as, resultantly, "disproportionately" strong I mean that they become stronger than one's desire to act from respect. In describing them as "judgment-resistant", I mean that they become altered, such that they can retain much of their strength, even after the moral law humiliates them.

I arrive at this definition of the tendency to evil: first, by looking at how Kant defines the propensity to evil in those passages where I take him to be talking about the propensity to evil *qua* tendency; second, by examining the kinds of immoral attitudes and action characteristic of the three grades of the drive to evil; and third, by investigating the difference between the evil and the good *Gesinnung*, and the relationship between one's tendency to evil and one's *Gesinnung*.[17]

The Tendency to Evil

Kant fairly clearly defines the propensity to evil in the following passage:

> by the concept of a propensity is understood a subjective determining ground of the power of choice *that precedes every deed*, and hence is itself not yet a *deed*. … Now, the term "deed" can in general apply just as well to the use of freedom through which the supreme maxim (either in favor of, or against, the law) is adopted in the power of choice, as to the use by which the actions themselves (materially considered, i.e. as regards the objects of the power of choice) are performed in accordance with that maxim. The propensity to evil is a deed in the first meaning (original sin), and at the same time the formal ground of every deed contrary to law according to the second meaning, [i.e. of a

> deed] that resists the law materially and is then called vice (derivative sin); and the first indebtedness remains even though the second may be repeatedly avoided (because of incentives that are not part of the law). (*Rel*, 6:31)

According to this passage, the propensity to evil is both the free decision to adopt an evil maxim as one's supreme maxim, and also "the formal ground of every {evil} deed". Indeed, a supreme, evil maxim just *is* the formal ground of every evil deed, i.e., what makes it possible for one to commit evil deeds.

Notice that Kant says that a person becomes capable of acting evilly only after she freely adopts an evil supreme maxim. I take the moment when someone freely adopts an evil supreme maxim to be identical to her choice to activate her propensity to evil,[18] and "the formal ground of every deed contrary to" the moral law to be the tendency to evil. Thus, not only does this passage spell out for us how a tendency to evil should be understood (as that which makes it possible for us to act immorally), but it also provides indirect support for my distinction between the susceptibility and the tendency to evil.

This is not the only passage where Kant elaborates on the propensity to evil, particularly in (what I take to be) its role as the tendency to evil. For example, he says that the propensity to evil "must consist in maxims of the power of choice contrary to the law" (*Rel*, 6:32).[19] Again, since this portrays the propensity to evil as a kind of drive or inclination—namely, the drive or inclination to adopt immoral maxims—this seems to refer to the tendency to evil.

Finally, it is important to take note of Kant's definition of propensity when he initially discusses the propensity to evil. "*Propensity* is actually only the *predisposition* to desire an enjoyment which, when the subject has experienced it, arouses *inclination* to it". On this definition, someone has a propensity to evil if and only if an inclination to

evil is aroused in her when she first experiences evil. I take this inclination to evil to be the tendency to evil.

Kant defines an inclination as a habitual desire.[20] It follows from this definition, and my identification of the tendency to evil with an inclination to evil, that the tendency to evil is a habitual desire for evil. However, as I showed in chapter 2, Kant understands a desire to be an attempted action.[21] Thus, to say that someone has an inclination for φ-ing is to say that she regularly attempts to φ. One explanation, though, for *why* someone regularly attempts to φ is that opportunities to φ present her with incentives—i.e., she is such that she wants to φ whenever she can. So, to say that someone has an inclination to evil is the same as saying that immoral actions present her with incentives or, what is the same thing, that she has a standing desire to engage in various kinds of immoral activity.[22]

We should note, though, the example Kant uses to clarify this definition of propensity, viz., that of savages' propensity for intoxicants. Savages who have a propensity for intoxicants *suddenly* develop *powerful* desires to consume intoxicants upon experiencing them. By the same token, someone with a propensity to evil is someone who suddenly generates powerful evil desires as soon as she experiences evil (this is why one need experience evil only once to activate one's tendency to evil).

On this reading, saying that someone has a tendency to evil amounts to claiming that she has powerful desires to perform immoral actions. Note that it is not saying that she *judges* immoral actions to be good (i.e. her desires for immorality do not necessarily lead her to adopt a motive in favor of them). Thus, merely having a tendency to evil, for all I have shown so far, is not the same thing as accepting a disposition-maxim according

to which one ought to engage in immoral actions or according to which engaging in immoral actions is good.

The Three Grades of the Propensity to Evil

Kant indirectly illuminates the tendency to evil by discussing the possible ways in which it can manifest itself through the three different "grades" of the propensity to evil, viz., frailty, impurity, and depravity.[23] It is worth exploring frailty, impurity, and depravity in detail, as each kind of immorality involves the propensity to evil. Importantly, since each grade of the propensity to evil has to do with immoral action, and since on my view immoral action (except for the choice to activate the propensity to evil) cannot take place unless the tendency to evil grounds it, it follows that the three grades of the propensity to evil are specifically grades of the propensity to evil *qua* tendency, so examining the three grades of evil—i.e. the kinds of immorality the tendency to evil allows for—gives us better understanding of the nature of the tendency to evil.

I display frailty when "I incorporate the good (the law) into the maxim of my power of choice; but this good, which is an irresistible incentive objectively or ideally, is subjectively the weaker (in comparison with inclination) whenever the maxim is to be followed" (*Rel*, 6:29). To put it in the terms I offered in chapter 2, someone shows frailty when she accepts both the Moral Maxim and another, immoral disposition-maxim and so experiences conflict. That is, someone who is frail judges actions in conformity with the moral law to be good, and desires them because of their goodness, but also wants, more strongly, something else that is immoral, and this stronger desire leads her to judge its object as good (because of the pleasure it will bring her). When such a person actually acts contrary to the moral law, then she acts from frailty.[24]

Someone acts impurely when her "actions conforming to duty are not done purely from duty" (*Rel*, 6:30), i.e., when she performs an action *A* in conformity with the moral law both because she accepts the Moral Maxim but also because she accepts another disposition-maxim according to which she should do *A* because it will make her happy.[25] By accompanying respect for the moral law with pleasure, though, impurity threatens to undercut the moral value of a person's actions.

For example, imagine someone begs me to lie while being cross-examined in court, but I refrain from doing it, not only because I think it is wrong to lie, but also because it pleases me to exercise power over someone's fate. This is dangerous, because the cooperating incentive can potentially muscle out respect as time goes on. I might initially refrain from lying both because it is wrong and because it makes me feel important, but eventually, I might do it only because it makes me feel important. This can result if I do not try to keep the desire to feel important in its proper place. Impurity, then, is dangerous because it weakens the strength of my moral motivation, eventually to the point where I will not do a moral action unless it also makes me happy. The impure person's *Willkür* "has not, as it should be [the case], adopted the law *alone* as its *sufficient* incentive but, on the contrary, often (and perhaps always) needs still other incentives besides it in order to determine the power of choice for what duty requires" (*Rel*, 6:30).

Finally, there is depravity. Kant defines depravity as:

> the propensity of the power of choice to maxims that subordinate the incentives of the moral law to others (not moral ones). … it reverses the ethical order as regards the incentives of a *free* power of choice; and although with this reversal there can still be legally good actions, yet the mind's attitude is thereby corrupted at its root (so far as the moral disposition is concerned), and hence the human being is designated as evil. (*Rel*, 6:30)

A person shows depravity when she subordinates morality to non-moral, i.e. sensible, desires. A person is depraved not merely when she wants to satisfy her sensible desires more than her moral obligations, for that is true of the frail person too. Rather, the depraved person judges that the satisfaction of at least some sensible desires is *overall* better than the discharging of her moral obligations; this is what Kant means when he says that depravity "reverses the ethical order as regards the incentives of a *free* power of choice".[26] Furthermore, this explains why someone who acts depravedly has a mind whose "attitude is … corrupted at its root (so far as the moral *Gesinnung* is concerned)". Kant adds that with depravity "the propensity to evil is … established (as regards actions) in the human being, even the best" (*Rel*, 6:30).

In section 3.1 I claimed that the tendency to evil gives a person desires to perform immoral actions, but not judgments of those actions' goodness (though I did not rule this out). I think investigation of the three grades of evil confirms this earlier definition but also allows us to expand it.

Take frailty first. On one construal, someone shows frailty when she merely wants to do something immoral more than she wants to do her duty. If this were all there were to frailty, then it would be perfectly in line with the 3.1 definition of the tendency to evil; indeed, it would just *be* the tendency to evil.[27]

However, according to the most common interpretation of frailty, and the one I share,[28] the frail person not only wants to satisfy illicit desires more than she wants to do her duty, she also sometimes, against her better judgment, acts on her illicit desires. Indeed, it is *only* when she acts akratically that she *expresses* frailty. Given the Maxim-Action thesis of the last chapter, according to which a person cannot act without also

acting on a disposition-maxim, and given as well that disposition-maxims consist of both desires and motives, it follows that someone who acts frailly not only desires to act immorally more strongly than she desires to act morally, she also judges that her immoral action is good (because it will bring her pleasure).

This does not mean that the tendency to evil creates a judgment that sensible desires are good. Rather, because everyone accepts the Prudential Maxim,[29] people almost always[30] judge their sensible desires to be good.[31] It follows that having a tendency to evil makes one more likely to act on certain of one's prudential maxims. But it is not as though the tendency to evil makes all of a person's sensible desires stronger (if that is what the tendency to evil did, we would not be able to know this, for Kant holds that all people in history (so far) have tendencies to evil—it would be like trying to spot the difference between the universe as it is now and the universe where everything is twice as large); rather, it makes just those desires that are immoral stronger. This raises two questions: how does the tendency to evil make desires stronger, and how does it "know" to make only the immoral desires stronger?

My speculation is this: the tendency to evil makes immoral desires stronger by making them judgment-resistant. To understand what I mean, we have to reexamine how the moral law provides incentives.

Normally, when a person judges that a sensible desire is immoral, she feels pain for entertaining the desire, and the desire is proportionately weakened. In addition to feeling pain, though, the person also feels respect – the fact that a mere judgment of something's immorality can have an effect on the strength of sensible desires awes her,

and makes her want to carry out the moral law's commands. In other words, a judgment gives rise to a desire.[32]

The tendency to evil is the flip side of respect. The tendency to evil fortifies desires against the moral law's humiliating power (this is not to say that the tendency to evil renders them completely impervious), and in so doing reduces the amount of respect for the moral law a person would otherwise feel.

If the tendency to evil fortified all immoral desires, it would not be noticeable; after all, the moral law does not humiliate desires in general, it humiliates desires that conflict with it. Thus, for a person to have a sense that the tendency to evil is working within her, it must be the case that it fortifies some immoral desires and not others. (Few, if any, people are immoral in every way; instead, they have particular vices.)

Because the tendency to evil affects only some sensible desires and not others, it follows that each of us is familiar both with what the moral law and with what the tendency to evil can do. We respect the moral law because we know it can weaken sensible desires that, when they do not go against it, are otherwise stronger. Similarly, on my view we have a kind of reverse-respect for the tendency to evil: because we know what the moral law can do, on those occasions when the tendency to evil buttresses our sensible desires against it, it enlivens them in the doing. In other words, by making a sensible desire judgment-resistant, it vivifies it.[33]

This answers the first question about how the tendency to evil makes desires stronger. But we still have the second question: given that it does not strengthen all sensible desires, how does it select which desires to strengthen? In brief: it depends on how one is constituted, what environment one finds oneself in, and what choices one

makes. I spell out this story in detail in chapter 5, sections 5 and 6. Now that we are armed with an understanding of the workings of the tendency to evil, we can look for greater confirmation of it as Kant's actual account by investigating the remaining grades of the drive to evil.

A person acts impurely when she does her duty both because it is her duty and also because she things doing so will give her pleasure. By the action-theory of chapter 2, a person's acting impurely means that she accepts two disposition-maxims: a moral maxim, "in circumstances *C*, doing action *A* is good (because it is morally required)", and a prudential maxim, "in *C*, doing *A* is good (because it brings pleasure)". By accepting the prudential maxim, though, she opens herself up to the possibility of failing to do *A*, even when it is morally required, because it will not bring about pleasure.

With impurity, the tendency to evil is, again, not responsible for a person's accepting a prudential maxim in the first place. Given everyone's allegiance to the Prudential Maxim, it follows that each of us always incorporates any desire into a disposition-maxim as soon as she has it. Moreover, there is nothing intrinsically wrong with wanting to do something both because it is morally required and because you expect to get pleasure from it. The tendency to evil, though, moves a person to retain her prudential disposition-maxim even after it is decoupled from her moral disposition-maxim.

Suppose, for example, I judge that I ought to help people, both because it is morally required (a moral disposition-maxim) and because it will improve my reputation (a prudential disposition-maxim). Should I encounter a situation where helping someone will not help my reputation, then the two disposition-maxims will decouple: I will judge,

on the one hand, that helping the person in this situation is good because it is morally required, but on the other hand I will see helping her as bad, because it will not improve my reputation. In such a case, the tendency to evil moves me to retain the prudential disposition-maxim by keeping the desire it has as an element at the same level of strength it had before, making it more difficult to refrain from acting on the (now, immoral) prudential disposition-maxim, and indeed enlivening it by making it resist the moral law's condemnations. This is why, with impurity, the moral disposition-maxim a person has is often not sufficient for action by itself; because, once the moral disposition-maxim in question decouples from its formerly partnered prudential disposition-maxim, it ends up appearing much the worse for the loss, thus inclining a person to subordinate the moral disposition-maxim to the prudential one.

With frailty and impurity, a person's tendency to evil makes some of her sensible desires judgment-resistant, thus bolstering their strength. The tendency to evil protects one's sensible desires from being weakened by as much as one's judgment of moral goodness is supposed to weaken them.[34] With depravity, though, one's sensible desires are not only stronger than one's moral desires; one also judges them to be *ultima facie* better than acting on one's moral desires. How the tendency to evil leads to depravity is a story I tell in section 4 of chapter 5. Once a person is depraved, though, the tendency to evil plays only a minor role. Since such a person already judges some kind or kinds of desire to be overall more worth acting on than one's duty, the tendency to evil no longer needs to defend against the moral law.

Gesinnungen

In some passages, Kant appears to define the propensity to evil as an evil *Gesinnung*. This cannot be right, though, for while he is clear that no one can ever extirpate her propensity to evil, he is equally insistent that each of us can revolt against her evil *Gesinnung* and replace it with a good *Gesinnung* (similarly, one's good *Gesinnung* is always in danger of being usurped by one's evil *Gesinnung*). In this section I show that the evil *Gesinnung* and the propensity to evil *qua* tendency are not identical.

While it is important to distinguish between the evil *Gesinnung* and the propensity to evil,[35] I explore *Gesinnungen* mainly to elucidate the tendency to evil. Articulating the relationship between *Gesinnungen* and the tendency to evil reveals once again that the tendency to evil is something in a person that causes her sensible desires to be judgment-resistant and so disproportionately strong.

3.3.1. The Evil *Gesinnung* Is Not the Propensity to Evil

Before elucidating the evil *Gesinnung* in detail, I shall offer my main reason for thinking it different from the propensity to evil *qua* tendency. It is this: Kant is clear that one cannot "extirpate" one's drive to evil, but he is equally insistent that one can "overcome" one's evil *Gesinnung*. It is, of course, possible to hold that extirpation is different from overcoming, such that overcoming one's evil *Gesinnung* does not imply anything about adopting a good *Gesinnung*. Inspection of the texts will show, though, that one overcomes one's evil *Gesinnung* only by adopting a good *Gesinnung*. I shall begin by looking at the passages in which Kant discusses the propensity to evil's ineradicability, and then show what it means to overcome an evil *Gesinnung*.

3.3.1.1. The Ineradicability of the Propensity to Evil

Kant speaks of the ineradicability of the propensity to evil *qua* tendency in three places:

> [I1] the use of freedom through which the supreme maxim (either in favor of, or against, the law) is adopted in the power of choice … {i.e.,} the formal ground of every deed contrary to law … is said to be … innate, because it cannot be eradicated (for the supreme maxim for that would have to be the maxim of the good, whereas in this propensity the maxim has been assumed to be evil). (*Rel*, 6:31-32)

> [I2] there is in the human being a natural propensity to evil … This evil is *radical*, since it corrupts the ground of all maxims; as natural propensity, it is also not to be *extirpated* through human forces, for this could only happen through good maxims – something that cannot take place if the subjective supreme ground of all maxims is presupposed to be corrupted. Yet it must equally be possible to *overcome* this evil, for it is found in the human being as acting freely. (*Rel*, 6:37)

> [I3] We cannot start out in the ethical training of our connatural moral predisposition to the good with an innocence which is natural to us but must rather begin from the presupposition of a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition; and, since the propensity to this [depravity] is inextirpable, with unremitting counteraction against it. Since this only leads to a progression from bad to better extending to infinity, it follows that the transformation of the *Gesinnung* of an evil human being into the *Gesinnung* of a good human being is to be posited in the change of the supreme inner ground of the adoption of all the human being's maxims in accordance with the ethical law, so far as this new ground (the new heart) is itself now unchangeable. (*Rel*, 6:51)

[I1] to [I3] display without question Kant's belief that the propensity to evil is inextirpable. The passages, though, do raise two questions. First, how can we be sure that Kant is talking about the propensity to evil *qua tendency* as opposed to *qua susceptibility*? If he meant the latter, then it could be that the tendency to evil was indeed the same as the evil *Gesinnung*, thus undermining the thesis of this (and the next) chapter. Second, while these passages indeed show the propensity to evil (in whatever sense) to be permanently present in each of us, they also give the impression that the propensity to evil is identical to the evil *Gesinnung*; so how does one who wants to separate the two dispel this impression?

I shall address the first question first. In [I1] Kant is clearly talking about the propensity to evil *qua* tendency because he claims that "the formal ground of every deed

contrary to law" is what is ineradicable. But this is much closer to what the tendency to evil (a judgment-insensitivity that can be applied to any of our sensible desires) is than to what the susceptibility to evil (an innate susceptibility to developing a drive to evil, given the meeting of the right triggering condition) is.[36]

In [I2] the "natural propensity" that is "not to be *extirpated* through human forces" is something that "corrupts the ground of all maxims". Again, insofar as this is something in us that affects the adoption of *every one* of our maxims, it fits much better with the tendency, rather than with the susceptibility, to evil.[37]

Finally, in [I3] the "propensity to this [depravity]" that is "inextirpable" is "a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition". Again, since this is a propensity to a depravity in our *Willkür* for adopting immoral disposition-maxims, the best understanding of it is as a tendency to evil.

We can now get to the second problem: does it not look as though Kant is himself equating the propensity to evil with the evil *Gesinnung* in each of these passages? To start, it will be helpful to have a rough-and-ready sense of what an evil *Gesinnung* is.

A *Gesinnung* is one's supreme maxim, i.e., a person's decision to make one of her highest maxims, either the Moral or the Prudential Maxim, authoritative over the other. One has an evil *Gesinnung* when one chooses to make one's obedience to the Moral Maxim conditional (in some sense) on one's satisfaction of the Prudential Maxim.

Now, in [I1] Kant claims that "the use of freedom through which the supreme maxim … is adopted in the power of choice … is said to be … innate, because it cannot be eradicated". In other words, the supreme maxim—one's *Gesinnung*—cannot be

eradicated. This certainly makes it look as though an evil *Gesinnung* is the propensity to evil. One must note, though, *why* one's evil *Gesinnung* cannot be eradicated: because "the supreme maxim for that would have to be the maxim of the good, whereas in this propensity the maxim has been assumed to be evil".

Let us parse this independent clause a bit. We cannot eradicate an evil *Gesinnung* because such eradication would require a supreme maxim that was good—i.e., to eradicate an evil *Gesinnung*, one would have to have a good *Gesinnung*. But why cannot one have a good *Gesinnung*? Because "in this propensity the {*Gesinnung*} has been assumed to be evil".

Here is how I, in light of my theory that the evil *Gesinnung* and the propensity to evil are not the same, read this clause. When one activates one's susceptibility to evil by adopting an evil *Gesinnung*, one not only makes the Prudential Maxim one's supreme maxim, one also brings upon oneself a tendency to evil that corrupts the grounds of one's lower disposition-maxims. It is not the evil supreme maxim that cannot be eradicated, but rather the supreme maxim "adopted in the power of choice"; the supreme maxim adopted in the power of choice, though, is not only the subordination of the Moral Maxim to the Prudential Maxim, *it is also the tendency to evil*. So it is the tendency to evil that cannot be eradicated.

What of Kant's remark that "the supreme maxim for {eradication} would have to be the maxim of the good, whereas in this propensity the maxim has been assumed to be evil"? It take it that he is saying the following: the tendency to evil is inextirpable *for us* because *we started out by adopting an evil* Gesinnung; however, if we, *counterfactually*, had adopted a good *Gesinnung*, then we would never have activated our susceptibility to

evil, and so would never have generated a tendency to evil. That is the only way the tendency to evil can be "eradicated"; it can be eradicated only if never activated in the first place.

The same approach we used in [I1] can be taken with [I2]. Kant there reasons that the propensity to evil cannot be extirpated because such an extirpation "could only happen through good maxims – something that cannot take place if the subjective supreme ground of all maxims is presupposed to be corrupted." Once again, the reason we cannot eliminate the propensity to evil is that we adopted an evil *Gesinnung* in the first place; this is why we even have a propensity to evil. If we had not made such a decision, then we would not have it; but given that we chose to activate our susceptibility to evil in the first place, we are permanently stuck with the results of our decision.[38]

In [I3] Kant is clearest about the distinction between the propensity to evil and the evil *Gesinnung*. He there claims that the "propensity to {adopting maxims contrary to the original ethical predisposition} is inextirpable", and so concludes that we must "unremittingly" struggle against it. Now, if the propensity to evil is literally inextirpable, then any struggle, even unremitting struggle, seems doomed to failure (in Kant's words, such a struggle would lead "only … to a progression from bad to better extending to infinity", but never to perfection). And yet despite the fact that we can never completely eliminate our propensity to evil, Kant is still confident that "the transformation of the *Gesinnung* of an evil human being into the *Gesinnung* of a good human being" can occur, for there can be a "change of the supreme inner ground of the adoption of all the human being's maxims in accordance with the ethical law", even though this change is

unobservable. Thus, we can never completely escape our propensity to evil, but we can transform our evil *Gesinnung* into a good *Gesinnung*.

3.3.1.2. Revolution in *Gesinnung*

Kant speaks about the transformation of an evil into a good *Gesinnung* in a number of places, but most important is what he says here:

> that a human being should become not merely *legally* good, but *morally* good (pleasing to God) i.e. virtuous according to the intelligible character [of virtue] (*virtus noumenon*)[39] and thus in need of no other incentive to recognize a duty except the representation of duty itself – that, so long as the foundation of the maxims of the human being remains impure, cannot be effected through gradual *reform* but must rather be effected through a *revolution* in the disposition {*Gesinnung*} of the human being (a transition to the maxim of holiness of disposition). And so a "new man" can come about only through a kind of rebirth, as it were a new creation … and a change of heart. (*Rel*, 6:47)

Here Kant says that a human being can become morally good (virtuous according to her intelligible character) only with a revolution in her *Gesinnung*. Assuming that this revolution is a change in *Gesinnung* from evil to good, it makes sense why it would have to be a sudden revolution rather than a gradual reform: because there is no middle ground between having an evil and a good *Gesinnung*.[40] One either makes the Moral Maxim supreme, or one does not (in which case one makes the Prudential Maxim supreme): "a human being {cannot} be morally good in some parts, and at the same time evil in others" (*Rel*, 6:24). For example, someone who always does her duty, except in regard to one kind of thing—say, her eating habits—has not really adopted the Moral Maxim as her supreme maxim (it would be akin to someone who thinks of herself as being Christian, except on Tuesdays – she would have to have some reason for why it is permissible for her to give up her Christianity one day a week, and any such reason will have to be one that subordinates Christianity to something else. Having such a reason, though, would mean she is not a Christian[41]).

Although a person can undergo a *Gesinnungsrevolution*, the result is not perfect conformity to the moral law. A "human being, who incorporates this purity {of the moral law} into his maxim, though on this account still not holy as such (for between maxim and deed there still is a wide gap), is nonetheless upon the road of endless progress toward holiness" (*Rel*, 6:46-47). Kant says here that someone who has incorporated the purity of the moral law into her maxims—who has restored "the original predisposition to good in" her—is a morally good person, but not necessarily someone who always acts in a morally good fashion. If the propensity to evil were the evil *Gesinnung*, then it is hard to see how to interpret this. Since, though, the propensity to evil *qua* tendency is different from the evil *Gesinnung*, and instead remains even after the evil *Gesinnung* is overthrown, this passage makes sense.[42]

3.3.2. The Evil *Gesinnung*

Everyone accepts two highest disposition-maxims: the Moral Maxim ("in any circumstances, I ought to do my duty") and the Prudential Maxim ("in any circumstances, I ought to do what makes me happy") or, as Kant calls it in the *Religion*, the law of self-love. The human being "incorporates the moral law into {his} maxims, together with the law of self-love" (*Rel*, 6:36). At times, though, the action that each highest disposition-maxim recommends or requires in some circumstances *C* will be different; on such occasions, a person has to decide which action to perform. If a person ends up willingly performing an action she knows the moral law to condemn, then she shows herself to regard the Prudential Maxim as supreme:

> In order … to call a human being evil, it must be possible to infer *a priori* from a number of consciously evil actions, or even from a single one, an underlying evil maxim, and, from this, the presence in the subject of a common ground, itself a maxim, of all particular morally evil maxims. (*Rel*, 6:20)

Which *Gesinnung* a person has depends on which of the two highest disposition-maxims she takes to be supreme. If she makes the Prudential Maxim supreme—if, that is, she judges that making herself happy is more important than doing her duty—then she has an evil *Gesinnung* and is an evil person. On the other hand, if she subordinates the Prudential Maxim to the Moral Maxim, then she has a good *Gesinnung* and is a good person:

> the difference, whether the human being is good or evil, must not lie in the difference between the incentives that he incorporates into his maxim (not in the material of the maxim) but in their *subordination* (in the form of the maxim): *which of the two he makes the condition of the other*. It follows that the human being (even the best) is evil only because he reverses the moral order of his incentives in incorporating them into his maxims. (*Rel*, 6:36)

Elsewhere, Kant writes that the *Gesinnung* is "the first subjective ground of the adoption of … maxims" (*Rel*, 6:25). If the character of a person's *Gesinnung* depends on which highest disposition-maxim she elevates above the other, then it makes sense why it should count as the first ground of all other maxims (i.e., her supreme disposition-maxim)—because her *Gesinnung* influences every one of her decisions.

Given that a person's *Gesinnung* influences every one of her decisions, one naturally concludes that the moral alignment of one's *Gesinnung* determines whether one is a good or evil person. We need to discuss, though, precisely how the *Gesinnung* affects a person's decisions, which discussion will also fully flesh out the tendency to evil.

Kant illuminates the evil *Gesinnung* in his discussion of depravity. Depravity "is the propensity of the power of choice to maxims that subordinate the incentives of the moral law to others (not moral ones)" (*Rel*, 6:30). Whereas frailty and impurity affect only particular disposition-maxims (i.e. a person could be frail in regard to her compliance with one disposition-maxim and impure in regard to another),[43] depravity

seems to affect the whole *Willkür*; someone is depraved if her *Willkür* has a propensity for adopting immoral maxims. But this makes depravity seem identical to the propensity to evil itself; what is the propensity to evil if not a propensity to adopting immoral maxims? Indeed, Kant even writes, "It will be noted that the propensity to evil is {with depravity} established (as regards actions) in the human being" (*Rel*, 6:30).

I do not think, though, that depravity is itself identical to the propensity to evil (*qua* tendency). One reason for thinking this is that depravity is supposed to be a grade of the propensity to evil; if depravity simply were the propensity to evil, it would be odd of Kant to say that it is also a grade of itself.

A better reason for distinguishing depravity from the propensity to evil *qua* tendency can be found in the latter part of Kant's description of depravity; depravity can:

> be called the *perversity* … of the human heart, for it reverses the ethical order as regards the incentives of a *free* power of choice; and although with this reversal there can still be legally good … actions, yet the mind's attitude is thereby corrupted at its root (so far as the moral *Gesinnung* is concerned), and hence the human being is designated as evil (*Rel*, 6:30).

Anyone who is depraved shows herself to have a mind whose attitude is "corrupted at its root"—i.e., anyone who is depraved shows herself to have an evil *Gesinnung*, and so "is designated as evil".

As was already shown in section 3.3.1, the propensity to evil *qua* tendency is different from the evil *Gesinnung*. Consequently, it is possible to have a good *Gesinnung*—i.e., to be good—while still having a tendency to evil. Since anyone who is depraved is evil, i.e., has an evil *Gesinnung*, it follows that depravity cannot be identical to the drive to evil.

What, then, is depravity? Like frailty and impurity, depravity shows itself in particular disposition-maxims. A person who acts frailly in regard to one kind of action

and impurely in regard to another kind may act depravedly with regard to a third kind. Acting depravedly in regard to a particular action, though, means reversing "the ethical order as regards the incentives of a *free*" *Willkür*. To act depravedly, my action must first of all be immoral, and second, I must also regard the prudential disposition-maxim of my action as better to act on than the moral disposition-maxim that the moral law suggests to me in its stead. Anyone who accepts the Moral Maxim will always judge any moral disposition-maxim as better than any prudential disposition-maxim, even if she does not always act on moral disposition-maxims. To judge a prudential disposition-maxim as better than a moral one, then, requires one to accept the Prudential Maxim as one's supreme maxim or, what is the same thing, to have an evil *Gesinnung*. This is not to say that someone who has an evil *Gesinnung* sees *every* kind of immoral action as better than every kind of moral action; but to judge even one kind of immoral action as *ultima facie* better than one kind of moral action entails having an evil *Gesinnung*. This is why anyone who acts depravedly, i.e. anyone who fully endorses acting on an immoral disposition-maxim, shows herself to have a mind "corrupted at its root", i.e. to be "radically" evil.[44]

What of Kant's claims that depravity "is the propensity of the power of choice to maxims that subordinate the incentives of the moral law to others (not moral ones)" and that "the propensity to evil is {with depravity} established (as regards actions) in the human being"? To understand these passages, we have to explicate the relationship between having an evil *Gesinnung* and having a tendency to evil.

3.3.3. The Evil *Gesinnung* and the Tendency to Evil

We know that the evil *Gesinnung* and the propensity to evil *qua* tendency are not identical. We also know depravity to entail an evil *Gesinnung* and *vice versa*. So we can

conclude that depravity is not identical to the drive to evil. But I would like to do more than merely conclude this; I would like to make sense of the troublesome passages that ended the previous section.

On my interpretation, each of us activates her susceptibility to evil, thereby generating a tendency to evil, as soon as she becomes rational. The act of bringing upon oneself a tendency to evil, though, is itself an act that can only be described as depraved. Thus, the act of triggering one's own susceptibility to evil expresses one's commitment to an evil *Gesinnung*. Once we realize this relationship between the evil *Gesinnung* and the initiation of the tendency to evil, we can explain away the appearance of conflation (on Kant's part) of the propensity to evil with depravity.

The tendency to evil comes into being only when someone adopts an evil *Gesinnung*. Thus, it is the product of a free choice, and so is something for which a person can be held responsible (indeed, this is why Kant routinely describes it as evil; see *Rel*, 6:29 and 37). We find Kant talking this way about both the propensity to evil *qua* tendency and the evil *Gesinnung*. Here is Kant on the *Gesinnung*:

> to have the one or the other *Gesinnung* by nature as an innate characteristic does not mean here that the *Gesinnung* has not been earned by the human being who harbors it, i.e. that he is not its author, but means rather that it has not been earned in time … This *Gesinnung* … must be adopted through the free power of choice, for otherwise it could not be imputed. (*Rel*, 6:25).

The *Gesinnung* is both innate and something we bring upon ourselves. Given that the evil *Gesinnung* differs from the propensity to evil, it follows that in describing the *Gesinnung* as innate, Kant is not merely saying, in different words, that the propensity to evil is innate. He is saying that the *Gesinnung* is innate, not because it is the propensity to evil, but because we have to assume that everyone starts out with a *Gesinnung*, in point of fact, an evil *Gesinnung*.[45] (While it is perhaps surprising that everyone starts out with an evil

Gesinnung, it is a matter of necessity that everyone starts out with one *Gesinnung* or the other; for reasons I spell out in chapter 2, section 3.2, Kant thinks everyone is committed to the Moral and the Prudential Maxims. As soon as a person has to choose between acting on either the Moral or the Prudential Maxim, she expresses her commitment to the supremacy of one or the other of these highest maxims, which commitment constitutes a *Gesinnung*.)

Just as the evil *Gesinnung* "is in fact grounded in freedom" (*Rel*, 6:25), so too the propensity to evil "can … be thought of … (if evil) as *brought* by the human being *upon* himself" (*Rel*, 6:29); indeed, it "must … always come about through one's own fault" (*Rel*, 6:32). Given the difference between the evil *Gesinnung* and the drive to evil, the best interpretation of Kant's remarks in 6:29 and 32 is that the propensity to evil *qua* tendency is activated by one's adoption of an evil *Gesinnung*.

Both the evil *Gesinnung* and the tendency to evil are the products of freedom—the first being the disposition-maxim of a free choice, the second a result of that choice. Additionally, though, the evil *Gesinnung* and the propensity to evil have the same scope: they both apply to all actions. The evil *Gesinnung* "applies to the entire use of freedom universally" (*Rel* 6:25) and the propensity to evil *qua* tendency is a "subjective determining ground of the power of choice *that precedes every deed*" (*Rel*, 6:32). From the fact that they have the same scope, though, one cannot draw the conclusion that they are the same thing. After all, both the Moral and the Prudential Maxim govern all maxims, but they are not the same thing.

The tendency to evil and the evil *Gesinnung* have the same scope but rule differently. The tendency to evil causes particular sensible desires to become resistant to

the diminishing effects of the deliverances of the moral law, while the evil *Gesinnung* makes action on certain sensible desires appear worthier than the discharging of certain of one's moral obligations. Admittedly, in applying to particular sensible desires neither the propensity to evil nor the evil *Gesinnung* regulate absolutely every sensible desire; for any sensible desire a person has, though, it could be affected by either the evil *Gesinnung* or the drive to evil. As mentioned in section 3.2 of this chapter, which sensible desires they regulate as a matter of fact is a matter of one's constitution and the process one underwent in becoming a depraved individual; but both govern *potentially* all sensible desires.

We are now positioned to assess Kant's claims about depravity with which I ended the last section, viz., that depravity "is the propensity of the power of choice to maxims that subordinate the incentives of the moral law to others (not moral ones)" and that "the propensity to evil is {with depravity} established (as regards actions) in the human being".

In the previous section I noted that depravity, like frailty and impurity, shows itself in individual disposition-maxims. But whereas it is not apt to call a person as a whole frail or impure, calling a person depraved is more fitting, since exhibiting depravity in even one disposition-maxim is enough to reveal a person to have an evil *Gesinnung*. Thus, "she is depraved" and "she has an evil *Gesinnung*" are, more or less, interchangeable. If we grant this, then saying that depravity is the propensity of the *Willkür* to immoral maxims is identical to saying that an evil *Gesinnung* is the propensity of the *Willkür* to immoral maxims. And this is true, for someone with an evil *Gesinnung* will adopt immoral maxims.[46]

Kant's claim that with depravity "the propensity to evil is … established (as regards actions) in the human being", is also now made intelligible. It could be that a person's becoming depraved—i.e. her taking on an evil *Gesinnung*—establishes the propensity to evil *qua* tendency. More likely, though, I think it means that with depravity the tendency to evil comes into full flower. The tendency to evil, after all, is a tendency to something; this something, I think, is becoming an evil person—making some sensible desire so judgment-resistant that it becomes judgment-proof. Once a person reaches this point, it is hard to imagine how she can escape. As Kant puts it, "How it is possible that a naturally evil human being should make himself into a good human being surpasses every concept of ours" (*Rel*, 6:44-45).

4. The Activation of the Susceptibility to Evil

I will examine this process of entering (and, to a lesser extent, leaving) depravity in chapter 5. Now, though, there is an odd and an end to be picked up: my claim that the triggering condition for the susceptibility to evil is being in an evil society; and, relatedly, my claim that there is a state of affairs in which a person does not activate her susceptibility to evil, namely, a world that exemplifies the highest good.

4.1. How the Susceptibility Is Activated

I claim that the condition that triggers the generation of the tendency to evil by the susceptibility to evil is living in an evil society. I do not want to give the wrong impression, though; it is not as though coming to maturity in an evil society literally *causes* a person to activate her idle propensity to evil. If that were the case, then society, rather than individuals, would be responsible for the triggering of the susceptibility to

evil. Societies do not cause the activation of the propensity to evil *qua* susceptibility, though; people freely activate their own propensities in response to social exemplars.

There are a couple of reasons for interpreting Kant in this way. First, there is the definition of a propensity itself. If someone has a propensity for *x*, then she will develop an inclination to *x* if she is exposed to *x* in the right way. So, and as noted at the beginning of section 3, someone who has a propensity to evil will develop an inclination to evil if she is exposed to evil in the right way.

Evil, though, is different from, say, alcohol. It is possible for someone to be causally responsible for the activation of her propensity for alcohol without being morally responsible. For example, she could unknowingly pick up a bottle of liquor and drink from it; someone could even force her to have some. In the first case, she would be causally but not morally responsible for her alcoholism; in the second case, she would be neither causally nor morally responsible for it. Evil, though, is not something that someone could accidentally engage in, or perpetrate by being coerced into it.[47] Instead, to do evil a person has to willingly go against what the moral law commits her to. So, for one to activate her susceptibility to evil/adopt an evil *Gesinnung* in a way for which she is morally blameworthy, it seems that she has to already have activated it/had an evil *Gesinnung*. This is why Kant writes:

> The rational origin … of … this propensity to evil, remains inexplicable to us, for, since it must itself be imputed to us, this supreme ground of all maxims must in turn require the adoption of an evil maxim. Evil can have originated only from moral evil. (*Rel*, 6:43)

Answering this problem is not on my agenda right now.[48] However, even though Kant thought that individuals are responsible for the activation of their own susceptibilities to evil, he also thought they needed to be in a community that tempted them to engage in it. Kant suggestively writes:

> The Scriptures express this incomprehensibility in a historical narrative … the human being … is represented as having lapsed into {evil} only *through temptation*, hence not as corrupted *fundamentally* (in his very first predisposition to the good) but, on the contrary, as still capable of improvement. (*Rel*, 6:43-44)

More definitively, Kant writes:

> Envy, addiction to power, avarice, and the malignant inclinations associated with these, assail his nature, which on its own is undemanding, *as soon as he is among human beings*. Nor is it necessary to assume that these are sunk into evil and are examples that lead him astray: it suffices that they are there, that they surround him, and that they are human beings, and they will mutually corrupt each other's moral *Gesinnung* and make one another evil. (*Rel*, 6:93-94)

This passage confirms that social influence is a precondition of people activating their own susceptibilities to evil. However, it seems also to confirm that it does not matter whether that society is evil or not ("Nor is it necessary to assume that these are sunk into evil and are examples that lead him astray").

While Kant does not say that the society has to be evil for people to activate their propensities to evil, he does assert that a propensity to evil can gain no foothold in a society that is good. He reasons that, even if a person undergoes a moral revolution, he is still subject to the evil *Gesinnung*'s usurpation:

> If no means could be found to establish a union which has for its end the prevention of this evil and the promotion of the good in the human being … however much the individual human being might do to escape from the dominion of this evil, he would still be held in incessant danger of relapsing into it. (*Rel*, 6:94)

How, then, can a person prevent his evil *Gesinnung* from regaining the helm? The only way is through the establishment of "an enduring and ever expanding society, solely designed for the preservation of morality by counteracting evil with united forces … only in this way can we hope for a victory of the good principle over the evil one" (*Rel*, 6:94).

If an evil society is a necessary condition for a person's generating a tendency to evil, and the susceptibility to evil cannot be activated in a good society, what does Kant mean when he claims that an evil society is not a precondition of a person's susceptibility

to evil giving rise to a tendency to evil? I suspect Kant was thinking of the most primitive society possible—one that barely qualifies as a society. Such a society is not good, nor is it evil; instead, it is innocent. Kant writes:

> It is not the instigation of nature that arouses what should properly be called the *passions*, which wreak such great devastation in {man's} originally good predisposition. His needs are but limited, and his state of mind in providing for them moderate and tranquil. (*Rel*, 6:93)[49]

Innocence, though, can be lost, as Kant is fond of reminding us: "There is something splendid about innocence; but what is bad about it, in turn, is that it cannot protect itself very well and is easily seduced" (*G*, 4:404-5). It does not take much for the innocents in such a primitive society to become guilty; indeed, all it takes is exposure to each other – their innate propensities to evil do the rest of the trick.

4.2. How to Prevent the Susceptibility's Activation

The highest good (or as I shall sometimes call it, "the best world" (*CPrR*, 5:125)) describes a state of affairs where everyone is perfectly virtuous and happy.[50] I claim that no one born into best world would activate her propensity to evil. The previous section gave some evidence for this proposition, for in it I quoted Kant as saying that in a world where everyone is unified to prevent evil and promote good—which qualification the best world would certainly meet—no one would have to worry about the reactivation of her evil *Gesinnung*. The main difficulty in asserting that a tendency to evil could not arise in the highest good is its relevance; the highest good might be nothing more than a motivating fiction, in which case we have no reason to think that the propensity to evil *qua* tendency is contingent. But if the drive to evil is not contingent, then it seems that Kant would have to use a transcendental argument to establish the UT; and I am skeptical

that one can be had. So we must now investigate whether the highest good could be realized.

In the *Critique of Pure Reason*, Kant refers to the highest good as an "ideal" (A810/B838), and in Mrongovius's lecture notes, Kant is reported to have said that "All ideals are fictions" (*LE-M*, 29:605). Why think ideals are fictional? Because "An ideal is the representation of a single thing, in which we depict … an Idea to ourselves *in concreto*" (*LE-M*, 29:605). However, we can never actually cognize any idea;[51] consequently, ideals are ways we make ideas vivid to ourselves in our imaginations.[52] Thus, while we can imagine an ideal, we can never experience one. Since the highest good is an ideal, it follows that we can never have an experience of the highest good; instead, it is a representation of the goal of the moral life.[53]

Ideas are fictions, while ideals are representations of ideas. Thus, it is not quite right to call ideals fictions. Instead, they are coherent imaginings, albeit ones that we could (perhaps) never be sure we could experience.[54] Indeed, Kant is clear that the best world is possible, albeit only in a practical sense of possible: "Only with this subordination {of happiness to virtue} is the *highest good* the whole object of pure practical reason, which must necessarily represent it as possible since it commands us to contribute everything possible to its production" (*CPrR*, 5:119). In other words, because morality commands that we do everything we can to bring about the best world, we are committed, for moral reasons, to believing that it is possible;[55] if theoretical reason convinced us that the highest good could not come about, then morality would be a sham.[56] However, there is nothing self-contradictory about the concept of a best world—although such a world coming about on its own would be so unlikely as to be almost

impossible—, so theoretical reason cannot undermine our belief in the possibility of the highest good. We are thus not only entitled to believe that the highest good can be realized, but obligated to try to bring it about.

We thus have reason to believe that the tendency to evil is contingent. Now the question is, why should we believe that everyone has a tendency to evil? This is the task of the next chapter.

---

[1] I coin these names in section 3.1 of chapter 2.
[2] See B3-4.
[3] Note that inclinations are drives in virtue of which things of a certain kind will give sensible incentives to an agent (or, what is the same thing, they are drives that give rise to desires for certain kinds of things). For more on inclinations, see section 2.2.3.1 of chapter 2.
[4] See *LA*, 25:1111-12, "Propensity ... is the inner possibility of an inclination, i.e. the natural predisposition to the inclination" (25:1111-2). See also *LA*, 25:1339 and 25:1517.The translations of all these passages can be found in Frierson 2005, 22.
[5] I derive this definition of propensities from the "simple conditional analysis" of Fara 2006. Whereas that analysis has to do with surefire dispositions (it reads: "An object is disposed to *M* when *C* iff it would *M* if it were the case that *C*" (Fara 2006, 4)), mine involves probabilistic dispositions (for the distinction between surefire and non-surefire dispositions, see Jansen 2007). Thus, applying Fara's analysis of dispositions to Kant's understanding of propensities gives us: "a person has a propensity for something *D* iff she is non-probabilistically disposed to form a disposition to *D* when put in circumstances *C*".
[6] I do not think we need to take literally Kant's claim that any savage who tries an intoxicant even once will *always* develop a nigh-inextinguishable desire for intoxication. Kant might have thought this, but nothing important seems to hang on it (except in the case of the propensity to evil, which is a deterministic, or surefire, second-order disposition).
[7] See *Ant*, 7:265.
[8] See *Rel*, 6:29 and 29n, where Kant offers a definition of propensity quite similar to the ones he presents in his anthropological works.
[9] See endnote 5.
[10] See section 2.2.3.1 of chapter 2 for a definition of K-desires.
[11] Actually, Kant writes of "predispositions to humanity" (*Rel*, 6:27).
[12] So, if it is the case that sociopaths cannot be moved by respect for the moral law, then they would not qualify as human *persons*, though they would be human *non*-persons.
[13] *Education* was compiled some time between 1777 and 1786, while the *Religion* was written in 1793, so there is no reason for thinking that Kant had yet categorized some predispositions as "original", with all that entails. Nonetheless, the faculties Kant presents in *Education* as needing development seem to be the same as the original predispositions he later describes in the *Religion*.
[14] Later in *Education*, Kant notes that education is supposed to supply four qualities to a person: "discipline", "culture", "discretion", and "moral training" (*E*, §18). Discipline is what allows a person to resist trying immediately to gratify her desires (*E*, §4), which shows that human beings are born not only with a predisposition to animality, but already with some, but only some, animal desires (e.g., the desire for sex does not emerge until the child is older). Thus, although the predisposition to animality starts out already activated, it does not begin fully developed. Culture provides a person with the ability to imagine more ends to pursue, and more ways to pursue them, than her animal desires alone ("It is culture which brings out ability. Ability is the possession of a faculty which is capable of being adapted to various ends" (*E*, §18)), while discretion informs a person how "to conduct himself in society, that he may be liked, and that he may gain influence" (*E*, §18). Culture and discretion together enable a person to imagine a

compossible set of desires, and so are necessary preconditions for a person to form a desire for happiness; so, the predisposition to humanity does not start out activated at all, but rather becomes so in the normal course of things. Finally, moral training trains a person so "that he shall choose none but good ends—good ends being those which are necessarily approved by everyone, and which may at the same time be the aim of everyone" (*E*, §18). This suggests that a person's desire to act on universalizable desires emerges—at least in its proper form—only under the right circumstances. Thus, the original predispositions, like the propensities, do not start out fully activated, but can only be activated as the human being matures.

[15] Note as well that predispositions have only to do with inclinations: "there is no question here of other predispositions except those that relate immediately to the faculty of desire and the exercise of the power of choice" (*Rel*, 6:28). So, a person's ability to, e.g., resist malaria because of sickle-cell anemia is not a contingent predisposition.

[16] Even fetuses eat, so even fetuses have predispositions to animality that are at least partially active. This is because the predisposition to animality "does not have reason at its root at all" (*Rel*, 6:28).

[17] There is another interpretation of what the tendency to evil is, according to which the tendency to evil makes us take certain of our sensible desires as providing us with *pro tanto* practical reasons, even when we judge that it would be immoral to act on them. I identify my reason for rejecting this interpretation in endnote 31.

[18] I explore this choice, which I take to be the adoption of an evil *Gesinnung*, in section 3.3.2 of this chapter.

[19] This is one of the passages that supports the reading of the tendency to evil as being that part of us in virtue of which we judge some of our sensible desires to be good, even when we also judge it to be immoral to act on them. See endnote 17.

[20] See *Rel*, 6:29, *MM*, 6:212 and *Ant*, 7:251 and 265.

[21] See section 2.2.3.1 and *Ant*, 7:251.

[22] Owing to the special nature of the tendency to evil, I think it is possible that it could qualify as an inclination to evil (i.e., as a habitual K-desire for evil) even if one never acts evilly (i.e., even if one never actually K-desires evil), excepting one's decision to activate the susceptibility to evil in the first place. What matters for Kant more than a person's actually acting evilly is the way she is prepared to act. Thus, if someone sees evil actions as sources of incentives, but owing to fortuity never has an opportunity to act immorally, then I think Kant would still count her as having an inclination (i.e., tendency) to evil, even though she never K-desired evil.

[23] See *Rel*, 6:29-30.

[24] It is not entirely clear that Kant thinks that frailty implies *acting* against the moral law (as opposed to merely having an immoral desire that stronger than one's moral desire), but his remark that "the frailty (*fragilitas*) of human nature is expressed even in the complaint of an Apostle: 'What I would, that I <u>do</u> not!'" (*Rel*, 6:29) suggests that he thinks people sometimes act immorally because of frailty.

[25] This disposition-maxim is of course ultimately grounded in the Prudential Maxim, but it need not be *immediately* grounded in the Prudential Maxim. I might, e.g., help my well-connected friend because I accept the disposition-maxim, "in circumstances where I have the opportunity, I ought to do what I can to advance my career", which maxim itself may be grounded in another disposition-maxim ("when I can do so without great cost, I ought to do what will give me respect from others"), and so on, until we reach the Prudential Maxim. See section 3.3 of chapter 2 for more on this.

[26] Kant also writes, "The will is depraved when <u>the motive power</u> of the understanding is outweighed by sensibility" (*LE-C*, 27:1429).

[27] Allison identifies frailty with the propensity to evil. See Allison 1990, 159.

[28] See endnote 24.

[29] See section 3.3 of chapter 2.

[30] I can conceive of one way in which one can have a sensible desire but not judge it good, despite one's adherence to the Prudential Maxim. One could experience a kind of malfunction, and have a desire for something that seemed to promise only pain (e.g., the desire to jump off a cliff or drive headlong into traffic). Because it offers only pain, one does not desire it because one expects happiness; and because acting on such a desire would be immoral, the moral law condemns it as well; while Kant never speaks of such desires, I see no reason to think that such desires, which resemble what is called an "urge" in Scanlon 1998, 38, could not arise, at least for a moment. Admittedly, an urge would be a desire not based in pleasure, and Kant claims that incentives are always grounded in the expectation of pleasure. But he also

claims that we cannot act on incentives without also incorporating them into disposition-maxims (the IT; see *Rel*, 6:23-24), and yet he allows for affects. On the same principle that, I imagine, moved Kant to make room for affects (namely, one should respect the appearances, and it sometimes appears as though people are controlled by their emotions), he should also allow for urges.

[31] I can now explain why we should not reduce the tendency to evil to the tendency to see immoral desires as presenting one with *pro tanto* reasons for action (see endnote 17). Adhering to the Prudential Maxim, as everyone does, means that everyone sees almost all of their sensible desires as giving them some reason to act, even if that reason is outweighed by morality. But surely the tendency to evil is not just one's adherence to the Prudential Maxim as one of one's highest disposition-maxims. (If it were, then Kant's denial of the tendency to evil in the 1780s (see chapter 4, sections 4 and 5) would make no sense.)

[32] See section 2.2.3.3 of chapter 2.

[33] I admit that Kant nowhere spells this out nearly as clearly as I would like him to; my support for it is not directly textual, but indirectly textual – I think this interpretation of the tendency to evil makes best sense of what Kant writes in the *Religion*. It gives the sense in which the propensity to evil suddenly generates a powerful inclination to evil when activated, and it also makes sense of why frailty operates in the way it does. I believe it also makes good sense of how impurity works (as for depravity, the fullest explanation of that occurs in chapter 5) and coheres well with Kant's conception of *Gesinnungen*.

[34] One's judgment that a sensible disposition-maxim would be morally wrong to act on must weaken the strength of the element of desire in that disposition-maxim to *some* degree; if the tendency to evil made the strength of one's sensible desires totally impervious to a judgment of their wrongness, then it would be impossible to have respect for the moral law, at least insofar as it required one to act differently from those sensible desires. This is because one's respect for the law stems from its power to weaken desires simply because of their nonconformity with it. Thus, if there were some nonconforming desires it could not weaken, then one would have no respect for it, at least with regard to its condemnation of those desires.

[35] Allison 2002 and Morgan 2005 both mistakenly identify the propensity to evil with the evil *Gesinnung*, causing problems for both of their interpretations.

[36] See also section 3.1 of this chapter.

[37] I should note that one scholar who equates the propensity to evil with the evil *Gesinnung* is John Hare, at Hare 1996, 53. But where Hare differs from people such as Allison, Card, and Morgan is in his attention to the fact that Kant writes we cannot extirpate the propensity to evil through *human* forces; i.e., we *can* extirpate the propensity to evil, but only through God's aid (see Hare 1996, 62-64). This certainly has more going for it than other views that conflate the propensity to evil with the evil *Gesinnung*, but I do not agree with it. However, I leave the presentation of my reasons for disagreeing with Hare for future work.

[38] I concede that this reading is strained, but at the end of the day I think that the evidence for identifying the propensity to evil with the evil *Gesinnung* (i.e., [I1] and [I2]) is weaker than the evidence against their identification, viz., the fact that the evil *Gesinnung* can be overcome, whereas the propensity to evil cannot.

[39] See also *Rel*, 6:14: "virtue, as a facility in *actions* conforming to duty (according to their legality), is called *virtus phaenomenon* but, as a constant *disposition* toward such actions from *duty* (because of their morality), is called *virtus noumenon*".

[40] This the thesis of "rigorism": "*The human being is* (by nature) *either morally good or morally evil*" (*Rel*, 6:22).

[41] I thank Patricia Kitcher for this example.

[42] See also *Rel*, 6:61-77, especially 68, where Kant reinforces the impression that one can have a good *Gesinnung* while also occasionally acting immorally.

[43] Some scholars (e.g. Allison 1990, 157-61, Card 2002, 76-77, and Sussman 2005, 159-62), seem to think that frailty and impurity affect a person as a whole. While there is certainly textual reason for thinking this (see *Rel*, 6:41-42), I think this cannot be right. For a person as a whole to be frail, she would have to be frail in regard to her *Gesinnung* – i.e. she would have to adopt the Moral Maxim as her supreme maxim, but occasionally adopt the Prudential Maxim as her supreme maxim. *Gesinnungen*, though, do not change willy-nilly like this. It makes even less sense to think impurity could apply to one's *Gesinnung*; if I adopt the Moral Maxim, but only insofar as it does not decouple from the Prudential Maxim, then I do not really adopt the Moral Maxim at all.

[44] See *Rel*, 6:32 and 38.

[45] "We cannot start out in the ethical training of our connatural moral predisposition to the good with an innocence which is natural to us but must rather begin from the presupposition of a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition" (*Rel*, 6:51).
[46] There is one sense, though, in which it is *not* true that the evil *Gesinnung* is the propensity to adopt immoral maxims; this is the sense that results when Kant uses "propensity" to refer to a disposition to give rise to inclinations given the right triggering conditions. Still, the evil *Gesinnung* is the triggering condition of the propensity to evil, so, while it is something of a stretch to say this is what Kant means when he says depravity is the propensity of the *Willkür* to adopt immoral maxims, I do not think it completely out of the realm of interpretive possibility.
[47] Obviously, someone could do something bad because forced to, but if she really had no choice in the matter, then what she did was not evil.
[48] I think there is a way Kant could have answered this problem, a way that he perhaps would not have endorsed but which does not alter much of his approach, and this is to adopt Strawson's "reactive attitude" approach. I intend to explore this possibility in future work.
[49] Admittedly, Kant's description here of pre-social humanity takes place in the midst of a *Gedankenexperiment* where someone who is "under attack" by his tendency to evil, and just sick about it, fantasizes that he is not to blame for his tendency's existence. In his reverie, he imagines that society must be to blame, for as a pre-social man he would have no cause to be evil. One might therefore take this dream to be unrepresentative of Kant's view.
[50] Kant claims that "happiness distributed in exact proportion to morality … constitutes the *highest good* of a possible world" (*CPrR*, 5:110-11). Now, Kant does not there explicitly say that in the best world happiness is proportioned in accordance to virtue *and everyone is perfectly virtuous*; for all he says, the best world could be one that contained only extremely unhappy evildoers. However, given that Kant says that striving to realize the highest good requires the postulation of an endless afterlife, so that one can completely conform one's will to the moral law (see *CPrR*, 5:122), it follows that he imagines a world exemplifying the highest good to be one where everyone is both perfectly happy and perfectly virtuous.
[51] See A320/B377 and *CJ*, 5:342.
[52] See *CJ*, 5:342 and *LPDR*, 28:994.
[53] See *CPrR*, 5:108 and *Rel*, 6:5.
[54] Why could the best world not be cognized? After all, a world where no one ever acts contrary to the moral law and is perfectly happy seems like a world we could experience. The problem, of course, is that you could never be sure of the perfect virtue of anyone in that world. As Kant says in the *Religion*, "we cannot observe maxims" (*Rel*, 6:20). However, he is also committed to the possibility of finite, holy beings "who could never be tempted to violate duty" (*MM*, 6:383); I should think beings with no tendencies to evil would qualify as such beings. If so, then we would have reason to think we could experience the best world—it would simply be a world where all of us were finite, holy beings. (But how would we know we were in a world like that? Presumably, the people in such a world would have no idea about a host of associated concepts: virtue, temptation, frailty, evil, etc. If no one could formulate any such idea, that would be strong evidence that we were in the best world.)
[55] See *CPrR*, 5:113.
[56] See *C2*, 5:114.

Chapter 4: Kant's Argument for the Universality Thesis

1. The Problem

In *Religion within the Boundaries of Mere Reason*, Kant claims everyone has a propensity to evil but notoriously adds that "We can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being, in view of the multitude of woeful examples that the experience of human *deeds* parades before us" (*Rel*, 6:32-33). This remark has puzzled interpreters for a variety of reasons.

First, most readers of Kant think that he equates the propensity to evil with the evil *Gesinnung*, which is a person's decision to rank what I call the Prudential Maxim ("in any circumstances, I ought to do what makes me happy") over the Moral Maxim ("in any circumstances, I ought to do my duty"). Since anyone who has an evil *Gesinnung* is an evil person, it seems that Kant is asserting that everyone is evil, but that the assertion needs no argument; instead, observation will do.

Second, interpreters generally assume that the propensity to evil is necessarily found in human beings; consequently, it seems that Kant is making the synthetic, *a priori* claim that every person, in every time and place, is evil. But then Kant's adverting to experience in justification of his claim is inexplicable; not only can observation not justify a synthetic, *a priori* claim, but that it cannot is a cornerstone of Kant's position in the *Critique of Pure Reason* that synthetic, *a priori* claims require transcendental arguments for their validation.[1] So Kant is the person we should *least* expect to find justifying a synthetic, *a priori* claim by recourse to mere observation; but there he goes.

Third, Kant says that we can "spare ourselves the formal proof" of the universality of the propensity to evil. His reference to a formal proof leads some to believe that he has a formal proof but that he is just sparing us its gory details. But Kant is not usually one to shy away from involved arguments. So why does he do so in this case?

In this chapter, I clarify why Kant gave no formal proof for the universality of the propensity to evil, and also provide the reasons that must have motivated him into accepting the universality of the propensity to evil in the first place. But before doing so, I need to clear up some misconceptions about what it is that Kant asserts.

2. A Proper Formulation of the Problem

I think the problem just described has been importantly misconceived by other interpreters. First of all, the propensity to evil and the evil *Gesinnung* cannot be the same, because whereas the propensity to evil is "inextirpable", the evil *Gesinnung* is something anyone can overcome.[2] Consequently, when Kant avers that we can spare ourselves the formal proof that everyone has a propensity to evil, he is not claiming that we can rely on observation to justify the claim that everyone is evil.

Second, the propensity to evil can be divided into two elements: the "tendency to evil", which is a property in someone that renders certain of her sensible desires resistant to the desire-debilitating judgments of the moral law;[3] and the "susceptibility to evil", which is simply the disposition to produce the tendency to evil in people given a particular triggering condition.[4] Because the propensity to evil *qua* susceptibility is a contingent predisposition, it follows that there is at least one circumstance in which it remains idle (Kant thinks this circumstance is a world that exemplifies the highest good).[5]

So, contrary to most Kant scholars, the propensity to evil (at least, *qua* tendency) is not necessary.

Now, when Kant says that we can rely on observation instead of a formal proof to show that the propensity to evil is universal, he means that we do not need to offer a transcendental argument for the claim that everyone who has yet existed has a propensity to evil *qua* tendency (I call the claim that everyone who has yet existed has or had a tendency to evil the "Universality Thesis" (UT)). But since in the highest good people do not have a drive to evil, it follows that he would not have been able to provide a formal proof anyway.

Finally, one must wonder why Kant did not explain why we need not provide a formal proof of the UT. Scholars who maintain that Kant needed to provide a formal proof, and who go about constructing the proof he needed, do not explain why, if he had available to him their favored proof, he did not offer it. No one has yet taken up this challenge.[6] For my part, I need to explain why, if Kant could not have given a formal proof of the UT, he did not simply say that.

3. The Story I Shall Tell

The questions we need to address, then, are these: (1) what was Kant's proof of the UT? (2) Why did Kant not offer this proof? To understand Kant's proof of the UT we first have to do some historical investigation. As it turns out, Kant denies the existence of a propensity to evil in three places in the 1780s: in the 1785-86 portion of his "Lectures on the Philosophical Doctrine of Religion", in the Collins manuscript of his lectures on ethics (which lectures inform his 1785 *Groundwork of the Metaphysics of Morals*), and in the 1777-86 *Education* (this last work, however, is not as important as the other two, for

in it Kant does not give any reason for denying the propensity to evil; consequently, I shall not analyze it).[7] Kant denies a propensity to evil in his lectures on religion because he cannot see how to square it with the omnibenevolence of God. The reason Kant has for denying a propensity to evil in his lectures on ethics is that such a propensity, if it existed, would make a hash out of his theory of unsociable sociability.

In the case of each denial Kant incurs a cost. In the lectures on religion, his denial of the propensity to evil ends up absolving not only God of responsibility for evil, but also us (he does not recognize this consequence in the course of outlining his theodicy, but he notes it in his 1791 "On the Miscarriage of All Philosophical Trials in Theodicy"). In the lectures on ethics, he recognizes that denying a propensity to evil forces him to deny the reality of something that appears to exist, viz., the devilish vices of envy, ingratitude, and *Schadenfreude*.

Because of the costs involved with denying a propensity to evil, Kant eventually does an about-face and not only admits (in 1791) that a propensity to evil exists, but (in 1793) incorporates it into his ethical philosophy as a fact that everyone must recognize. In particular, everyone must recognize that everyone else, including herself, has a propensity to evil *qua* tendency (i.e., the UT).

Kant's argument for the UT has to do with the costs that came from his original denial of the propensity to evil. His denial of the propensity to evil forced him to deny both our moral responsibility for evil and the reality of devilish vice, but asserting the drive to evil stems from our now-assumed culpability for immorality (should we commit any) and the reality of devilish vice. To oversimplify the argument somewhat: because everyone would be culpable for her immorality (should she engage in any) and because

everyone is capable of committing devilish vice, we can know that everyone has a tendency to evil. In other words, if you grant that people would be culpable for any immorality they engaged in, or are capable of devilish vice, then you are logically committed in each case to the claim that those people must have a tendency to evil, for neither culpability for immorality nor devilish vice is possible without it.

That is how Kant's argument for the UT works. Why does Kant not give it? I conjecture that his reasons for not giving it have to do with the considerations that show the UT to be correct to begin with. Let me clarify. In the part of the *Religion* before Kant, seemingly blithely, announces that he need not give a formal proof for the UT, Kant spells out his concept of the propensity to evil. Anyone who understands the propensity to evil, and accepts Kant's claims about it, will know that real immorality is not possible without it. Thus, one who accepts Kant's analysis of the UT, but who nonetheless denies that the propensity to evil *qua* tendency is universal, has to deny culpability for immorality or the reality of devilish vice. But if someone seriously denies either of these things, then she is arguing in bad faith; and there is no point in providing a proof for someone who argues in bad faith. Rather, you simply have to show her the evil that exists all around her. That will do the trick better than any argument.

4. Kant's Theodicy

In his "Lectures on the Philosophical Doctrine of Religion",[8] Kant claims that "Evil has no special germ" and "*A special germ toward evil cannot be thought*" (*LPDR*, 1078). Similarly, in the Collins manuscript of Kant's lectures on ethics (dating from 1784)[9] Kant is reported to have said, "in the nature of man's soul there resides no immediate inclination to evil" (*LE-C*, 27:440-41). However, the contexts of these two

quotations differ, and it is not obvious that Kant means to deny the existence of what he would later call the "propensity to evil". Thus, we need to take a closer look at them.

Kant clearly tries to give a theodicy in the "Philosophical Doctrine". For instance, he rhetorically asks, "Is evil … inevitable, and in such a way does God really will evil?" (*LPDR*, 28:1078) and answers, "Not at all; but rather God wills the *elimination* of evil *through the all-powerful development of the germ toward perfection*" (*LPDR*, 28:1078-79). Kant pretends to wonder whether God is to blame for the evil in the world, and to still his wondering produces a theodicy.[10]

To see how Kant's theodicy works, we first of all have to understand what he means by "evil". Kant seems to define evil as the subordination of moral obligations to sensuous desires. He writes, "the strength of {man's} instincts will beguile him and he will abandon himself to them, and *thus arises evil*" (*LPDR*, 27:1078); similarly, in the *Groundwork* he claims, "if I deviate from the principle of duty this is quite certainly evil" (G, 4:402-3). In other words, Kant uses "evil" primarily to refer to evil *actions* rather than evil *people*. With his theodicy, then, Kant wants to explain why God permits evil actions, i.e., why God allows people to subordinate their moral obligations to their sensuous desires.

Kant's reduction of evil to immoral actions has two consequences. First, pain is not itself an evil the existence of which needs to be explained; instead, Kant seems to see pain, at least pain that is not caused by a free being, as a morally neutral phenomenon. Pain is not evil; it is the pain free beings cause that is evil. Indeed, pain has negative moral weight only if we accord it too much importance, and suffer as a result: "all happiness or unhappiness depends on ourselves, and on the way our minds accept the

situation" (*LE-C*, 27:367).[11] So, while experiencing pain is not something for which people can be culpable, suffering – placing too much weight on one's pain – is something for which people can be held responsible.

Thus, not only does Kant face no problem of animal suffering, he does not need to justify God's permission of "natural evil" (the pain people experience from natural disasters and diseases) at all. Rather, all he needs to explain is why God allows people to act immorally (and why God allows people to suffer—why, that is, he allows them to react to their pain as negatively as they do).

Second, because only immoral actions (and suffering) need to be explained, evil can emerge only as soon as people are morally responsible, i.e., capable of recognizing and acting on their moral obligations. Thus Kant claims, "when the human being begins to use his reason, he falls into foolishness … *the first development of our reason toward the good is the origin of evil*" (*LPDR*, 28:1078) and "the human being … finds evil *first* when his reason has developed itself far enough that he recognizes his obligations" (*LPDR*, 28:1078-79).

Given how Kant defines evil, we should expect him to use a "free will theodicy",[12] i.e., a theodicy according to which God permits evil because he wants us to have free will, a good whose goodness outweighs the costs of its possible abuses. Indeed, not only is free will a good whose goodness outweighs any potential evil, free will is the *sine qua non* of goodness. Kant writes that animals':

> actions contain {animal necessity}. If all creatures had such a choice, tied to sensory drives, the world would have no value. But the inner worth of the world, the *summum bonum*, is freedom according to a choice that is not necessitated to act. Freedom is thus the inner worth of the world. (*LE-C*, 27:344)

If God had not gifted people with freedom, there would be no goodness in the world aside from that which resulted from his free actions.[13]

It is not obvious that Kant uses a free will theodicy, though, or at least not the typical one that sees evil as the result of immoral actions that are free in the libertarian sense. While Kant believes in libertarian free will, most of the time he seems to think that immoral decisions result simply from having an animal nature, rather than from the abuse of freedom. For example, he writes, "Evil is … not a means to good, but rather arises as a by-product, since the human being has to struggle with his own limits, with his animal instincts" (*LPDR*, 28:1078); along the same lines, he claims, "if we ask where the evil in individual human beings comes from, the answer is that it exists on account of the limits necessary to every creature. It is just as if we were to ask: Where do the parts of the whole come from?" (*LPDR*, 28:1079) Both explanations of evil see it as not just as a possible misuse of freedom, but as a *necessary* one. Kant's view appears to be that immoral actions—subordinations of morality to sensuous desires—will happen simply because a person has animal desires, because she is finite. I therefore call Kant's theodicy a "theodicy of finitude".

It is not obvious, though, why a free being who has animal desires will necessarily *succumb* to them. There seems nothing inconceivable about imagining someone who has animal desires and yet always resists them whenever they counsel deviation from morality; indeed, making oneself into such a being is precisely the goal the moral law gives us (*CPrR*, 5:122-24). What allows Kant to see a person's having an animal nature as necessarily resulting in her departing from the moral law is her uncultivatedness:

> Evil has no special germ; for it is mere negation and consists only in the limitation of the good. It is nothing beyond this, other than incompleteness in the development of the germ to the good out of uncultivatedness. The good, however, has a germ; for it is self-

> sufficient. This predisposition to good, which God has placed in the human being, must be developed by the human being himself before the good can make its appearance. But since at the same time the human being has many instincts belonging to animality, and since he has to have them if he is to continue being human, the strength of his instincts will beguile him and he will abandon himself to them, and thus arises evil, or rather, when the human being begins to use his reason, he falls into foolishness. (*LPDR*, 28:1078)

Each of us starts out with an undeveloped predisposition to the good, but a developed predisposition to animality. That is, we all have strong animal desires, but only an inkling of how to think morally (Kant does not elaborate on what this inkling is, but it is perhaps simply the ability to simulate others' thoughts and thereby imagine that they are like us, and so are as important as we are). As a result, we are much more likely to follow our animal desires than we are to treat people as ends in themselves. Indeed, it seems that to develop our moral thinking, we *must* first act immorally. You do not appreciate how important it is to act morally until you treat someone (or perhaps are yourself treated) as a mere means: "To begin his cultivation he must step forth out of his uncultivated state and free himself from his instincts. – But what then will be his lot? Only false steps and foolishness" (*LPDR*, 28:1077). Consequently, in order for a being with both an incompletely developed predisposition to the good and a well-developed predisposition to animality to appreciate the bindingness of the moral law, she has to first of all act immorally. This is why a free being with animal desires will necessarily succumb to them.

Evil thus results from the combination of an incompletely developed predisposition to the good and a more highly developed predisposition to animality. But insofar as evil results from failing to appreciate the bindingness of the moral law, it is hard to see how a person can be fully morally culpable for failing to adhere to the moral law. After all, when she acts immorally, she does not really know what she is getting into.

Kant could perhaps admit that the first departure from morality is not free, and instead focus on the fact that people who have very strong animal desires and little respect for morality are quite *likely* to act evilly. But insofar as the great strength of a person's animal desires overwhelm her and thus make her act immorally, her moral responsibility is once again diminished (although not eliminated).

Kant does not see things this way, though. He insists that although their strong animal desires increase the probability that people will depart from morality, it is still the case that their immorality is something they freely will:

> As soon as the human being recognizes his obligation to the good and yet does evil, then he is worthy of punishment, because he could have overcome his instincts. And even the instincts are placed in him for the good; but that he exaggerates them is his own fault, not God's. (*LPDR*, 28:1079)[14]

Kant, then, takes the following line: people do evil because they are born with strong animal desires and only a weak disposition towards moral action. Therefore, God is not responsible for the evil people do. Instead, the people themselves are responsible for their evil, because when they satisfy their illicit animal desires, they also know what they are doing is wrong, and moreover they can refrain. Thus, Kant appears to use a free will theodicy after all.

But this is puzzling, for on the one hand, Kant strongly implies that evil comes from the bare fact of people's having needs, rather than from the abuse of their free will:

> The possibility of deviating from the moral law must adhere to every creature. For it is unthinkable that any creature could be without needs and limits. God alone is without limitations. But if every creature has needs and deficiencies, then it must also be possible that impulses of sense (for these derive from the needs) can seduce it into forsaking morality. (*LPDR*, 28:1113)

Kant wants to claim that evil results just from our having needs, for this gets God entirely off the hook. Even if he had wanted to, God could not have created a being without

needs, for a being can be without needs only if it is infinite, and it is logically impossible for God, an infinite being, to create another infinite being.[15]

On the other hand, though, Kant wants to say that the evil we do is a result of our free choice, for once we recognize what our obligations are, we are capable of living up to them, and therefore culpable for not doing so. Kant's final position, then, is this: to excuse God from responsibility for evil, he blames evil on our finitude; but to ensure that we still have moral responsibility, he blames evil on our free decisions. The problem is, if evil is the result of finitude, it seems that it is not something for which we could be really responsible, for it was unavoidable; and if evil is something for which we are really responsible, it is not really the result of finitude, but the result of an abuse of free will.

Kant has the resources to construct an answer to this problem, although I do not think it ultimately works. As Henry Allison has pointed out, for Kant the necessary condition of an agent's action being free is that she agent-causes it; it does not matter whether she could have refrained from agent-causing it.[16] Thus, evil could stem both from our finitude and from our free will: because of our finitude, we must necessarily choose to depart from the moral law, but because our evil choices are agent-caused (albeit ones we must make) we make them freely.

Unfortunately, even if our immoral choices are free in this agent-causal sense, it does not follow that Kant can see them as ones for which we are morally responsible. If I *had* to agent-cause a certain decision (and we should rightly wonder how sensible desires could *necessitate* a free substance into causing itself to do something), then even if I freely make that decision (in some sense of "freely"), it is still the case that I am not morally responsible for it: "from the practical point of view this idea {of a prototype of

humanity pleasing to God}[17] has complete reality within itself. For it resides in our morally-legislative reason. We *ought* to conform to it, and therefore we must *be able* to" (*Rel*, 6:62). As Derk Pereboom notes:

> it seems reasonable to interpret Kant as supposing … that if "ought" principles are true or hold for us, it must in general be the case that we are able to act in accord with them. The following moral 'ought implies can' principle is attractive: If one ought to do something, then it must be the case that one can do it. (Pereboom 2006b, 559-60)

We are now in a position to see what Kant means when he denies the existence of a "special germ toward evil" or an "immediate inclination to evil". Immoral actions result merely from a person's having incompletely developed predispositions to humanity and personality; anyone in such a state will let herself be overcome by her sensuous desires, either because of her ignorance of morality or because she does not think herself strong enough to resist her desires. Given that evil results simply from uncultivatedness, there is no need to posit anything else to explain why people act immorally. As Kant puts it:

> evil in the world can be regarded as incompleteness in the development of the germ toward the good. Evil has no special germ; for it is mere negation and consists only in the limitation of the good. It is nothing beyond this, other than incompleteness in the development of the germ to the good out of uncultivatedness. (*LPDR*, 28:1078)

But a propensity to evil, as Kant defines it in the *Religion*, is something over and above uncultivatedness. As Kant makes clear, in what could be a conscious repudiation of his views in the lectures on religion, we cannot blame our animal desires for the evil we do:

> the ground of … evil cannot … be placed, as is commonly done, in the sensuous nature of the human being, and in the natural inclinations originating from it. For not only do these bear no direct relation to evil (they rather give the occasion for what the moral disposition can demonstrate in its power, for virtue): we also cannot presume ourselves responsible for their existence … though we can well be responsible for the propensity to evil which, since it concerns the morality of the subject and hence is to be found in the latter as a freely acting being, must be capable of being imputed to the subject as itself guilty of it (*Rel*, 6:34-35).

This passage shows not only that Kant's explanation of evil changed in the interim between his lectures on religion and the *Religion*, but also gives us some insight into why

he changed his mind – because blaming one's sensuous desires as the source of one's own evil eliminates one's own moral responsibility for evil.

The problem with the view expressed in the lectures on religion is that, while it does not negate human responsibility for evil, it does diminish it. In the *Religion*, Kant can say not only that evil comes about from weakness and ignorance—i.e. uncultivatedness—, but also that there is something in us that *wants* to give in to our desires at morality's expense. In other words, we sometimes perpetrate immorality fully willingly, because we see that immorality as better than what the moral law commands. It is because of this fundamental evil attitude, this evil *Gesinnung*, that we can be evil in the thoroughgoing sense familiar to us from experience. In contrast, our animal desires "bear no direct relation to evil".

5. Denying the Propensity to Evil As the Source of Devilish Vice

In the manuscript of Collins's 1784 notes on Kant's lectures on ethics, Collins reports Kant to have claimed that "There is reason to believe … that in the nature of man's soul there resides no immediate inclination to evil, but that its tendency is evil only in an indirect fashion" (*LE-C*, 27:440-41). However, Kant goes on to clarify his point with the remark, "Man … has no direct inclination towards evil *qua* evil, but only an indirect one" (*LE-C*, 27:441). Similarly, in the *Religion* Kant claims that there is no such thing as "diabolical" evil, i.e., "a disposition (a subjective *principle* of maxims) to incorporate evil *qua evil* for incentive into one's maxim" (*Rel*, 6:37). Thus, it may appear to some that when Kant denies the possibility of a direct or immediate inclination to evil, he is doing no more than denying the possibility of doing evil for evil's sake, i.e.,

diabolical evil. If that is correct, then nothing Kant says in the Collins manuscript shows him to deny a propensity to evil.

Let me quickly rebut this objection. As I will show in this section, Kant thinks that a corollary of denying the direct inclination to evil is a denial of the devilish vices. Yet in the *Religion*, the book in which Kant denies diabolical evil, he asserts the possibility of diabolical vice: "the vices that are grafted upon {the inclination to gain worth in the opinion of others} can … be named vices of *culture*, and in their extreme degree of malignancy (where they are simply the idea of a maximum of evil that surpasses humanity), e.g. in *envy*, *ingratitude*, *joy in others' misfortunes*, etc., they are called *diabolical vices*" (*Rel*, 6:27).[18] Consequently, it cannot be the case that Kant thinks that the devilish vices require diabolical evil for their possibility.[19]

Just because the direct inclination to evil is not diabolical evil, it does not follow that it is the propensity to evil either. And truth be told, it is not exactly the propensity to evil; but the same reason that moves Kant to reject the immediate inclination for evil also applies to the propensity to evil (in particular, *qua* tendency). So though Kant does not really reject the propensity to evil in the Collins lectures on ethics, he rejects something close enough.

The Devilish Vices and the Direct Inclination to Evil

Just before denying an immediate inclination to evil, Kant discusses the devilish vices. There are three: ingratitude, envy, and *Schadenfreude*.[20] In each of his descriptions of them, Kant compares what one could call the "full-blooded" version of the vice (i.e., the devilish version) with a lesser form (i.e., what I call the "natural" version, by which I

mean the version in keeping with nature's designs). Let us look at these three vices, in both their lesser and devilish forms, one at a time.

Kant sees run-of-the-mill ingratitude as resulting from a person's being ashamed at having to receive favors from others. "All men are ashamed at receiving favors, since they thereby incur obligations, and the other acquires calls and claims on the person he has shown favor to" (*LE-C*, 27:439). People do not like to receive favors (at least the ones they have to pay back) from others because falling into another's debt makes one vulnerable in the future. But Kant says more than just that people "do not like" to have others do them favors; they are *ashamed* to. Presumably, a person is ashamed to receive a favor from someone else because she thinks her dependence on him says something about who she is—her debt to him makes it more difficult for her to entertain as high an opinion of herself as she would like.[21] This is why Kant goes on to say "a strong-minded man will therefore not accept favors, in order not to be bound" (*LE-C*, 27:439).

The shame people can feel at being bound to another is key to the development of ingratitude. The more proud someone is, the more shame he will feel at being in another's debt. If he is proud enough, then he will refuse to recompense, or perhaps even admit that he had received, a favor. At this point, he displays ingratitude: "this is already a motive to ingratitude, if the recipient of favor be proud and selfish, since from pride he will feel shame at being beholden to the other; and from selfishness he will not concede his indebtedness, and so becomes defiant and ungrateful" (*LE-C*, 27:439). This is a natural level of ingratitude—refusing to admit that someone else has done you a favor, or failing to pay someone back from a sense of injured pride. However, natural ingratitude can increase to a devilish level: "If this ingratitude increases so much that he cannot

endure his benefactor, and becomes his enemy, that is the devilish degree of the vice, since it is utterly repugnant to human nature, to hate and persecute those who have done one a kindness" (*LE-C*, 27:439). In other words, someone is devilishly ungrateful if he tries to harm his benefactor simply because she has done him a favor; such ingratitude is "devilish" because "it is utterly repugnant to human nature" to try to damage those who have favored you simply because they have done you a favor (and, presumably, thereby made you feel inferior to them).

The next devilish vice is envy. Envy comes from "grudging", the displeasure a person feels in the greater happiness of others. Kant gives several examples of begrudging others their good fortune:

> if I am discontented, and every one else is in good spirits, then I grudge it to them … If I alone have poor fare to eat, and everyone else is faring well, that vexes me, and I grudge it them … Death can be borne, for all men must die; but were all to live, and I alone should have to die, that would vex me greatly. (*LE-C*, 27:438-39)

Kant seems to be of the view that *whenever* others are generally happier than you, you hold a grudge: "We take our stand on the relativities of things, not on the things themselves. We are grudging, because others are happier than we are" (*LE-C*, 27:439).[22] Because grudging results simply from perceiving you are particularly worse off than most (perhaps all) others, "Even good-natured souls feel grudging" (*LE-C*, 27:438); "There is thus a grudging element in our nature" (*LE-C*, 27:439).[23]

Kant writes as though there is a progression from grudging to envy (at the least, there is a common element of displeasure at others' good fortune). "We grudge, when displeased at another's advantage; we are too much put down by his good fortune, and therefore grudge it to him. But if we are displeased at the fact that the other has any share of happiness, that is envy" (*LE-C*, 27:438). Based on Kant's treatment of ingratitude, one

might expect that the difference between grudging and envy is that whereas the grudging person merely dislikes others' greater happiness, the envious person actually tries to harm others because they are happier. But this is not what we find; instead of talking about how the envious person acts differently from the grudging person, Kant discusses the different attitude of the envious: "envy is when we wish imperfection and ill-fortune to others, not so that we might ourselves be perfect or fortunate in consequence, but so that in that case we might alone be perfect and fortunate. The envious man … seeks the sweetness of happiness in this, that he alone enjoys it, and all others are unhappy" (*LE-C*, 27:438); and "the man of envy not only wishes for happiness, but wants it all for himself. He would like to enjoy it with misery all about him, and only so can he be fully content with his happiness" (*LE-C*, 27:440).

This is not how we commonly understand envy. The *Oxford English Dictionary* defines envy as "The feeling of mortification and ill-will occasioned by the contemplation of superior advantages possessed by another"[24]—a definition broad enough to include both grudging and Kant's "envy", but surely closer to how Kant understands grudging. The extremity of Kant's definition of envy brings up the question of how it could result from something as tame as grudging.

Someone grudges when her perception of others' greater happiness makes her receive an additional increment of unhappiness. By contrast, someone envies when she wishes others to become less happy, perhaps to the point that those others have no happiness at all, while her happiness remains constant or increases. Kant seems to distinguish two kinds of envy: an attitude where you want others, presumably those who are happier than you, to become less happy, and an attitude where you want to be the

only one who is at all happy. It is this second kind of envy that is the devilish version of envy: "The envious man wishes to be happy when all around him are unhappy, and seeks the sweetness of happiness in this, that he alone enjoys it, and all others are unhappy. This is the envy of which we will soon be learning that it is demonic" (*LE-C*, 27:438).[25]

Assuming there are these gradations in envy, we can see how one could go from grudge to envy to devilish envy. First, you begrudge some group of people their happiness—you resent their greater happiness (for example, because it makes you feel small). Your resentment toward them makes you envious—you want them to become less happy, and are pleased if something diminishes their fortunes. Finally, your envy can become devilish—given the inverse relationship between their unhappiness and your happiness, at the limit you would want them to become as unhappy as possible—that is, possessed of no happiness whatsoever—, so that you may become exultant. Indeed, the dependence of your happiness on their unhappiness might even move you to take action to bring this result about; the devilishly envious man "is … seeking to eradicate happiness throughout the world, and is thus an insufferable creature" (*LE-C*, 27:440).

The last devilish vice is *Schadenfreude* (sometimes translated as "malice" (*LE-C*, 27:380) and "malicious glee" (*LE-C*, 27:439)). Kant writes that *Schadenfreude* "consists in taking an immediate pleasure in the misfortunes of others; for example, by trying to stir up strife in a marriage, and suchlike, and then gloating at the parties' troubles" (*LE-C*, 27:440). As this definition makes clear, *Schadenfreude* consists of an attitude that can lead to immoral action; the *schadenfreudlich* person takes pleasure in at least some others' suffering, and because of this may try to cause suffering, so that she can enjoy it. Note that envy and *Schadenfreude* are different; the envious person takes pleasure in the

suffering only of a select group of people who are happier than she is, whereas the *schadenfreudlich* person may take pleasure in the suffering of a person not because he was or is happier than she, but just because he is suffering; this is true even if he is unhappier, and even if his happiness would not make the *schadenfreudlich* person feel small. As Kant notes, "*Schadenfreude* has a different complexion {from envy}. Such people laugh when others weep, and feel pleasure when others feel pain" (*LE-C*, 27:443).

Kant does not distinguish between natural and devilish versions of *Schadenfreude*. However, in the 1793 Vigilantius lecture notes, where Kant makes much clearer the distinction between natural and devilish versions of vices, he describes natural *Schadenfreude* as the kind where a person merely takes pleasure in the suffering of others, while devilish *Schadenfreude* is the kind that leads to immoral action.[26] Because his 1784 discussion of the devilish vices tends to mirror, in an inchoate way, his later 1793 and 1797 (*The Metaphysics of Morals*) discussions of them, there is some reason to think that this is how he would have distinguished between natural and devilish *Schadenfreude* in 1784.[27] Assuming this distinction is right, the naturally *schadenfreudlich* person may still have respect for others, even if she finds their suffering amusing (as long as the suffering she witnesses is not horrendous), while the person who would cause others to suffer just to enjoy herself shows herself to have a much more instrumental attitude towards people.

These, then, are the natural and devilish versions of the vices: someone shows natural ingratitude if she refuses to pay back or acknowledge a favor she received; she exhibits devilish ingratitude if she sees the fact that someone benefited her as a reason to harm him. A person is naturally envious if she wants others who are happier than she to

slip down a few pegs and fall at least to her level of happiness, if not a little farther; she is devilishly envious if she wants to be the only person who has any happiness at all. Finally, a person shows natural *Schadenfreude* if she takes pleasure in others' misfortunes, while she displays devilish *Schadenfreude* if she is willing to cause others' misfortunes just so she can enjoy them.

With an understanding of this context, we can now properly evaluate Kant's denial of a direct inclination to evil. Kant writes that:

> All three, ingratitude … envy and *Schadenfreude*, are devilish vices, because they evince an immediate inclination to evil. That man should have a mediate inclination to evil is human and natural; the miser, for example, would like to acquire everything; but he takes no pleasure in the other having nothing at all. (*LE-C*, 27:440)

The difference between a direct and an indirect ("mediate") inclination to evil is that one with an indirect inclination to evil acts immorally only as a means to some end that is not in itself immoral. The miser wants everything, and having an unlimited amount of stuff is not in itself evil. However, in his pursuit of things, he is liable to act immorally, which is evil. The devilishly envious person, on the other hand, not only wants as much for himself as he can imagine, he also takes an additional pleasure in his having gained through others' loss. That is, the bare fact that others lose their possessions is pleasing to him, independent of its connection to his gaining possessions. That is why devilish envy involves a direct inclination to evil—because the ends it moves people to have are themselves immoral, and their immorality is key to the envier wanting them.

This analysis applies not just to envy, but also to ingratitude and *Schadenfreude*. The naturally ungrateful person fails to repay her debts because she wants to maintain her self-esteem, which is a permissible end (although the action taken to achieve it is immoral). By contrast, the devilishly ungrateful person not only fails to repay her debts,

but she takes the fact that she is indebted to give her reason to hurt the person who helped her in the first place. She reasons this way because she has the affirmation of her superiority to others as her end, which is intrinsically impermissible.

As for *Schadenfreude*, natural *Schadenfreude* is a spontaneous reaction, and so is not connected with having an end one way or another. Devilish *Schadenfreude*, though, aims to produce occasions that trigger instances of *Schadenfreude*, which is an end, and an intrinsically immoral one at that.

Kant tells us that the three devilish vices evince a direct inclination to evil. After telling us this, though, he adds:

> There is reason to believe … that in the nature of man's soul there resides no immediate inclination to evil, but that its tendency is evil only in an indirect fashion. Man cannot be so ungrateful as actually to hate his benefactor; he is merely far too proud to be thankful to him, and for the rest, wishes him every happiness; the only thing is, he would like to be well out of his way. Nor does he have any immediate urge, either, to rejoice at another's misfortune, save only that if, for example, a person has come to grief, we are pleased because he was puffed-up, rich and selfish; for men would like to preserve equality. Man therefore has no direct inclination towards evil *qua* evil, but only an indirect one. Yet *Schadenfreude* is often already strongly apparent in the young. … It is, however, a sort of animality, whereby man retains something of the beast in him, which he cannot overcome. (*LE-C*, 27:440-41)

So after delineating the devilish vices, each of which expresses a direct inclination to evil, Kant turns around and denies the devilish vices. For instance, he denies that devilish ingratitude exists: "Man cannot be so ungrateful as actually to hate his benefactor; he is merely far too proud to be thankful to him … the only thing is, he would like to be well out of his way." In other words, although it appears as though agents are capable of taking others' charity as a reason to cause them harm, this appearance is deceiving. E.g., when I harm you, ostensibly because you benefit me, what is really going on is nothing more than natural ingratitude: I harm you, not because I want to assert your superiority, but instead because I have convinced myself that it is the best way to no longer have to

deal with my debt to you. I hope my injuring you persuades you to drop the debt, which allows me to maintain the self-esteem I had before. There are thus no actions expressing devilish vice.

Kant explains away devilish *Schadenfreude* in a similar manner. He asserts that we do not "have any immediate urge … to rejoice at another's misfortune, save only that if, for example, a person has come to grief, we are pleased because he was puffed-up, rich and selfish; for men would like to preserve equality."[28] We do not enjoy just anybody's misfortune; rather, we are made happier by the suffering of those we think deserve it. This seems to have to do only with natural *Schadenfreude*, not devilish *Schadenfreude*;[29] however, given that devilish *Schadenfreude* presupposes natural *Schadenfreude*, it follows that the devilishly *schadenfreudlich* person seeks to bring about suffering only for those people who think of themselves as superior in the first place, which does not evince a direct inclination to evil, but rather only a normal desire for equality.[30]

Thus, when Kant denies the existence of a direct inclination to evil, he denies that people can want to perform immoral actions to bring about ends that are also immoral (and are regarded as such by the performing agent). This is why he insists that devilishly vicious action has to be reinterpreted so that it counts only as an instance of normally vicious action. More important, this shows that when he rejects a direct inclination to evil, he is rejecting what he later came to call the propensity to evil (in particular, that part of the propensity to evil I call the tendency to evil); for a propensity to evil *qua* tendency, like a direct inclination to evil, allows us to think of ourselves as morally superior to others (at least, at the stage of depravity), and to perform immoral actions simply because they allow us to assert our moral superiority, an immoral end.[31]

The question all this brings up, though, is why Kant discussed the devilish vices in the first place, if his only intention was to deny that they were as he described them. The reason is that people appear to engage in devilishly vicious behavior; thus, fidelity to common experience demands they be discussed. But if devilish vices were as they appeared to be, then Kant would have to retool his theory of "unsociable sociability", his "favorite idea" (Wood 1999, 214). Thus, rather than tweak his theory of unsociable sociability, Kant decided to explain away the devilish vices. Obviously, though, I need to prove this.

The Devilish Vices and Unsociable Sociability

Before I can argue for the incompatibility of devilish vice and unsociable sociability, I must explain what unsociable sociability is. But first, some context.

In his "Idea for a Universal History from a Cosmopolitan Perspective", Kant sketches his theory of history. The premise of this theory is found in the "First Proposition", "*All of a creature's natural predispositions are destined eventually to develop fully and in accordance with their purpose*" (*IUH*, 8:18), or as he also puts it, "An organ that is not meant to be used, or an arrangement that does not achieve its purpose, is a contradiction in the teleological theory of nature" (*IUH*, 8:18). Thus, even though no one has fully developed predispositions to humanity and personality,[32] we know that they are supposed to become fully developed, for otherwise we would not have them. The problem is, they can become fully developed only in the human species as it perfects itself through time, not in individuals (*IUH*, 8:18-19). This, then, is the basis of Kant's theory of historical development: we must interpret the history of the human race as progressing in such a way that its predispositions gradually develop.

Kant thinks unsociable sociability is the vehicle of this development. He defines it as human beings' "tendency to enter into society, a tendency connected, however, with a constant resistance that continually threatens to break up this society" (*IUH*, 8:20). The idea is, people feel a need to be in the society of others, "because in such a condition they feel themselves to be more human, that is to say, more in a position to develop their natural predispositions" (*IUH*, 8:20-21). When in the company of others, though, they want to compete with and surpass their fellows, for in doing so they achieve an intoxicating sense of self-worth. Kant calls this desire an "unsociable trait that predisposes {people} to want to direct everything only to their own ends and hence to expect to encounter resistance everywhere, just as they know that they themselves tend to resist others" (*IUH*, 8:21). Their competition with others causes them not only to improve themselves, but also to improve the species, for the knowledge that people gain in self-development is passed on to their offspring and becomes embedded in social institutions. Kant avers that "Without those characteristics of unsociability … human beings would live the arcadian life of shepherds, in full harmony, contentment, and mutual love. But all human talents would thus lie eternally dormant" (*IUH*, 8:21).

Kant's theory of historical development offers an explanation of how the race has developed over time, and also a prediction of how it will evolve in the future. However, in order for Kant to be able to explain the behavior of humans, *en masse*, at all times, he has to have a theory of human nature, a guess at what it is about us that explains why we act in the way we do. And his theory of human nature involves the same explanatory principles as his theory of history: by nature we have certain undeveloped

predispositions, and by nature we are moved to interact in ways that result in the development of those predispositions.

That means, though, that if people act in ways that hinder the development of those predispositions, they are acting counter to human nature. And as it turns out, this is Kant's problem with the devilish vices. For example, Kant writes of ingratitude:

> If … ingratitude increases so much that he cannot endure his benefactor, and becomes his enemy, that is the devilish degree of the vice, since it is utterly repugnant to human nature, to hate and persecute those who have done one a kindness, and since it would also cause harm, if all men were thereby deterred from well-doing, and so became misanthropes, on seeing that they would be ill-used for their benevolence. (*LE-C*, 27:439-40)

Why is ingratitude of this level "devilish"? Kant offers two reasons: first, because "it is utterly repugnant to human nature", and second, because "it would also cause harm, if all men were thereby deterred from well-doing, and so became misanthropes, on seeing that they would be ill-used for their benevolence." With this second reason, Kant's language, which seems to invoke considerations of non-universalizability, indicates that in order for something to be a devilish vice, it has to be immoral.

But what of this first reason – how does ingratitude of the devilish degree go against human nature? Devilish ingratitude is contrary to human nature because it goes against the way nature intends to advance the species. Nature wants to use unsociable sociability to develop our propensities; in particular, nature makes our self-esteem depend on how skillful and on how much status we have relative to others. Making our self-esteem dependent in this way not only causes us to develop our predispositions—we develop them in order to become socially better-regarded or more skillful than others, so we can improve our self-esteem—, which is good for the species as a whole, but also encourages us to benefit one another with gifts. If I can give you a gift (thereby indebting

you), not only do I set myself up for recompense down the line, but I also increase my self-esteem—I show myself to be socially superior or more skillful than the one I benefit, which allows me to raise my self-estimation. So unsociable sociability—the societal delusion that self-esteem is a zero-sum game—brings with it acts of charity, which improves the species.

Now, simple ingratitude does not thwart this goal. If I benefit someone and she does not pay me back, I suffer to some degree, but I still have the psychic pleasure of knowing that I acted morally better than she on this occasion. Devilish ingratitude, though, thwarts this goal: the devilishly ungrateful person not only fails to return my favor, she uses it as an excuse to harm me. In most cases, the possible danger that devilish ingratitude threatens outweighs the possible psychic benefit it offers. Thus, devilish ingratitude is in tension with the mechanisms of unsociable sociability. It is for this reason, I think, that Kant regards it as contrary to human nature.

The block-quote from *LE-C*, 27:439-40 is not the only place where Kant describes a devilish vice as contrary to human nature. In one place, Kant labels the three devilish vices as "devilish" because "they evince an immediate inclination to evil." Importantly, he continues, "That man should have a mediate inclination to evil is human and natural" (*LE-C*, 27:440); the clear implication is that a person having an immediate inclination to evil is *inhuman* (i.e., devilish) and *unnatural*.

In another passage, Kant avers:

> In regard to his vices, man can go astray in two directions, that of baseness, or brutality, where by violating duties to his person, for example, he demeans himself below the beasts; and that of wickedness, or devilry, where a man makes it his business to pursue evil, so that no good inclination survives. So long as he retains a good disposition, and the wish to be good, he is still a man; but if he commits himself to wickedness, he becomes a devil. (*LE-C*, 27:464)

Again, Kant sees someone not in the thrall of the devilish vices as "still a man", but someone who "commits himself to wickedness" as having become "a devil."

Also of interest in this passage is Kant's comparison of the devilish to the bestial vices. On Kant's view, we can become inhuman in two ways: by making ourselves devils, and by making ourselves brutes. In his words:

> certain of man's vices are human, in that they accord with his nature, even though they are vices – for example, lying; but others are such that they lie outside humanity, and cannot be reconciled at all with the nature and character of man. Such vices are of two kinds, the beastly and the devilish. (*LE-C*, 27:380)

Just as there are three devilish vices, there are also three bestial vices: gluttony, drunkenness, and *crimina contra naturum* (*LE-C*, 27:380).

In contrast to his treatment of the devilish vices, Kant does not say that seeming instances of drunkenness, gluttony, and sexual perversity are not in fact what they appear. However, he does say they are contrary to nature—in particular, animal nature. In talking of sexual perversity, for example, Kant says:

> A *crimen carnis* is a misuse of the sexual impulse. Every use of it outside the state of wedlock is a misuse of it, or *crimen carnis*. All *crimina carnis* are either {either natural sexual crimes or unnatural sexual crimes}. The former are contrary to sound reason; the latter, to our animal nature. (*LE-C*, 27:390)

That is, a behavior is sexually perverse because it subverts the natural teleology of our sexual organs. Animals (Kant thinks) do not engage in such sexually perverse activity, because they always act to satisfy their naturally given ends. Consequently, only people can be sexually perverse, and one who misuses his sexual nature thus "debase{s} {his} human condition below that of the animal" (*LE-C*, 27:391).

So, people engage in bestial vice when they use their reason to act contrary to their animal desires' natural objects. By analogy, people engage in devilish vice when they use their reason to go against their non-animal desires' natural objects. Since there

are, broadly speaking, only two kinds of non-animal desires (socially developed desires and moral desires), it follows that someone acts devilishly viciously when she uses her reason to act against her social and moral desires' natural objects.

But there is more to devilish vice than just this; remember, in addition to devilish envy, ingratitude, and (possibly) *Schadenfreude*, there are natural versions of those vices. Both devilish and natural vice, though, go against social and moral desires. So what distinguishes them from each other?

Again, analogizing devilish to bestial desires illuminates the issue. When speaking of *crimina carnis*, he distinguishes between natural and unnatural sexual "crimes". Natural ones include premarital sex, adultery, and incest; they are in keeping with nature because they involve intercourse of two sexes, but they are contrary to reason because (presumably) maxims standing behind such actions cannot be universalized.[33] Unnatural sexual vices include masturbation, homosexuality, and bestiality. They are contrary to animal nature because they are not the kind of acts that can lead to offspring.[34]

The best way of applying this distinction to envy, ingratitude, and *Schadenfreude* has them divided into normal vices that are contrary to reason but not to nature's end for us—i.e., improvement of the species—and into devilish vices that are contrary, not only to reason, but also to nature's end for us. In other words, and as I have been arguing, what makes a vice devilish is its undermining nature's goal of improving the species, that is, of going against human nature.

Before going on, I should note something about what it means for a devilish vice to go against human nature. To go against human nature, it is not enough for a behavior

actually to undermine the development of a predisposition—after all, even given his re-description of devilish vices, they are still behaviors that undermine the development of our predispositions. What separates a natural from a devilish vice is that a devilish vice must *aim* at undermining human predispositions (not under this precise description, though). If such psychological states are possible (and with the depravity enabled by the tendency to evil, they would be), then nature has no guarantee that her goals will be met, because they will give rise to action often enough. However, if we rule out such mental states as impossible, then instances where we actually undermine the predispositions' development will be much less common.

6. Kant's Change of Mind on the Propensity to Evil

The first place where one can detect a change in Kant's thinking on evil and the devilish vices—and his invocation of a propensity to evil to explain them—is in his 1791 essay, "On the Miscarriage of All Philosophical Trials in Theodicy".

Kant's First Assertion of An Evil Propensity

Kant's overriding aim in "On the Miscarriage" is to show that no extant theodicy works, but also to show that the problem of evil fails as well.[35] The explanation for the failure of both is the same: we cannot know God's reasons for permitting evil, but similarly, we cannot know that God could not have good reasons for permitting evil. Consequently, we should not engage in the task of theodicy at all; instead, we should lead a morally upright life, and if we succeed, a faith in God will come that will withstand the doubts the problem of evil raises.[36]

I do not intend to provide a detailed exegesis of "On the Miscarriage". There are, however, three points from the essay I want to emphasize: Kant's definition of evil; his

rejection of his own earlier theodicy; and his assertion of an innate propensity to evil that is itself evil.

Kant presents three kinds of counterpurposiveness in his theodicy essay, though only the first two are relevant to my purposes. The first kind of counterpurposiveness is "The absolutely counterpurposive, or what cannot be condoned or desired either as end or means" (*OM*, 8:256). Kant also calls this first kind of counterpurposiveness the "morally counterpurposive" and "evil proper (sin)" (*OM*, 8:256-57). It is not surprising that sin cannot be condoned, either as an end or as a means; but it is puzzling that Kant describes sin as something that cannot even be *desired* as an end or a means, for after all, people sin all the time. The explanation for this is that Kant is talking about the condonability and desirability of sin from God's perspective (before describing the three kinds of counterpurposiveness, he writes, "Now whatever is counterpurposive in the world, and may be opposed to the wisdom of its creator, is of a threefold kind" (*OM*, 8:256)); thus, sin is behavior that cannot be condoned or desired by God, either as a means or as an end.

This makes sin contrast with the second kind of counterpurposiveness, "The conditionally counterpurposive, or what can indeed never co-exist with the wisdom of a will as end, yet can do so as means" (*OM*, 8:256). Another name for the conditionally counterpurposive is "ill (pain)" (*OM*, 8:256-57). Pain is only conditionally counterpurposive, because while God would never will it as an end to be realized, he can will it as a necessary means for securing some other end (say, the advancement of a finite agent's virtue).

What is important about sin is that it, unlike pain, is absolutely counterpurposive: we cannot imagine it advancing any of God's purposes by his allowing it, and yet, since it

clearly does exist, it makes us doubt God's holiness.[37] This is some confirmation that sin, at least sometimes, undercuts the benefits of unsociable sociability. Unsociable sociability, after all, is a way of showing how vices like (the non-devilish versions of) ingratitude, envy, and *Schadenfreude* fit into God's providential plan.[38] If this is right, and sin does not square with unsociable sociability, it follows that Kant at this point must countenance the possibility of the devilish vices that he denied five years earlier.

Another significant facet of sin is the role it plays in the problem of evil. Because sin is absolutely counterpurposive, it is not easy to imagine why God would permit it. However, theodicies have been proposed, for instance Leibniz's best-of-all-possible-worlds theodicy (which Kant says "can be freely given over to the detestation of every human being who has the least feeling for morality" (*OM*, 8:258)), as well as the free will theodicy.[39] More interestingly, though, Kant brings up and dismisses his own theodicy of finitude:

> The second alleged vindication [for God's permitting sin] would indeed allow for the actuality of moral evil in the world, but it would excuse the author of the world on the ground that it could not be prevented, because founded upon the limitations of the nature of human beings, as finite. (*OM*, 8:258-59)

The problem with the theodicy of finitude is the one identified earlier; while it gets God off the hook, it also eliminates our own moral responsibility for evil: "the evil would thereby be justified, and, since it could not be attributed to human beings as something for which they are to be blamed, we would have to cease calling it "a moral evil"" (*OM*, 8:259).[40]

This is an important remark, for two reasons. First, it decisively shows that Kant changed his mind about his earlier theodicy of finitude. Second, it shows *why* he changed his mind; before 1791, he understood evil as stemming from our sensible desires: we

commit evil when our desires overpower us.[41] If this is the case, though, then our desires, and not we, are responsible for the evil we do. This is why he thought of his earlier theodicy as undercutting moral responsibility for evil.[42]

Kant finds all the theodicies he lists (there are nine of them) to be unsatisfying, for different reasons. However, it also appears to be his view that the people who offer theodicies do so in bad faith. He uses the Book of Job to illustrate this point. Job, a morally righteous man, doubts God's goodness because of the intense misfortune he suffers. His comforters, presumably less righteous than he, offer a variety of reasons to convince him that his suffering is warranted. These reasons, like the theodicies Kant lists, fail to convince, because they are based on alleged theoretical insight into God's reasons. But we can have no theoretical insight into God's reasons, because they lie in the noumenal realm.

The reason our lack of (theoretical) insight into God's reasons is important is that it allows Kant to conclude that people who nonetheless claim to know God's reasons are not being sincere.[43] On Kant's view, we cannot know God's reasons for permitting moral evil, pain, and injustice (the third kind of counterpurposiveness).[44] And yet not only do numerous divines claim to know God's reasons, they claim to known them even though other divines claim a similar knowledge, and yet forward different reasons, as Kant's nine different theodicies illustrate.[45] It is this fact that moves Kant to claim (speaking of Job's comforters) that they "speak as if they were being secretly listened to by the mighty one, over whose cause they are passing judgment, and as if gaining his favor through their judgment were closer to their heart than the truth" (*OM*, 8:265).

Worse than the insincerity motivating theodicies is the institution of public statements of faith. When priests make the asseveration of religious creeds mandatory under threat of punishment, they help to perpetuate a community-wide insincerity: "these blind and external *professions* (which can very easily be reconciled with an internal profession just as false) can, if they yield *means of gain*, bring about a certain falsehood in a community's very way of thinking" (*OM*, 8:269). The problem is, sincerity (or as Kant also call it, "formal conscientiousness" (*OM*, 8:268))—saying and judging as true only those things one believes to be true[46]—is the very minimum required for good moral character.[47] And yet, even though it is a baseline condition for good character, we still admire it greatly, which is evidence that it is rare and to be prized. Assuming that sincerity is both rare and the minimum that can be expected of a good person, it follows that there must be an innate propensity for mendacity in people:

> honesty (mere simplicity and straightforwardness of mind) is the least that we can possibly require of a good character (especially if we waive candor of heart) and it is therefore difficult to see on what that admiration which we reserve for such a character is based; it must be that sincerity is the property farthest removed from human nature … None but a contemplative misanthrope … can hesitate whether to find human beings to *deserve hatred* or rather *contempt*. The properties for which he would judge them qualified for the first finding are those through which they do deliberate harm. That property, however, which appears to him to expose them to the second estimate, could be none other than a propensity which is *in itself evil* even if it harms no one – a propensity for something which cannot be used as means for any purpose; something which, objectively, is good in no respect. The first evil would indeed be none other than the evil of *hostility* … the second can be none other than *mendacity* (*OM*, 8:270).

This propensity for mendacity motivates certain evils. For example, it produces egregious theodicies in the first place (those who formulate them do not honestly advance them, but instead hope to flatter God; without this propensity for mendacity, people would not take flattering God to have more going for it than being sincere), and moreover it explains why coerced public statements of faith, of which such theodicies are often part, can hold such sway without provoking mass outrage.

In addition to producing evil, the propensity for mendacity is also itself evil.[48] First, it enables self-deception, which, if adopted as a principle, undercuts the possibility of developing good moral character at all (which character enables the adoption of the moral law as one's maxim; see *Ant*, 7:294):

> The evil of {the propensity for mendacity} is baseness, whereby all character is denied to the human being. – I am here restricting myself principally to the impurity that lies deep in what is hidden, where the human being knows how to distort even inner declarations before his own conscience. (*OM*, 8:270)

Second, though, the propensity for mendacity is evil because it serves no social function, unlike the propensity for hostility:

> The *first* inclination {i.e. the propensity for hostility} has a purpose whose function is yet permissible and good in certain farther connections, e.g. hostility against incorrigible disturbers of the peace. The *second* propensity {for mendacity}, however, is to use a means (the lie) which is good in no respect, whatever its aim, since it is evil and reprehensible in itself. The *evil* with which competence for good ends in certain external relations can yet be associated is in the constitution of a human being of the first kind; it is a sinning in means, which are not, however, reprehensible in every respect. (*OM*, 8:270)

The propensity for mendacity, unlike the propensity for hostility, is evil because there is no good it serves by its existence. This strongly suggests that the propensity for mendacity competes against the goods brought about through unsociable sociability, which are enabled by the propensity for hostility.

Although the propensity for mendacity is not the tendency to evil (the former has to do simply with lying, whether to ourselves or others, whereas the latter gradually moves a person to act depravedly in all manner of ways), it is clearly an anticipation of the propensity to evil *qua* tendency in several respects. First, it is itself evil. Second, it is woven into human nature. And third, it enables evils that do not themselves promote any social good. Since "On the Miscarriage" is where Kant repudiates his earlier theodicy, introduces us to the notion of moral evil as absolutely counterpurposive, and also introduces the notion of an innate, evil propensity for mendacity, we have reason to

suspect that all these changes are in some way related. My suggestion is that absolutely counterpurposive evil is what motivates Kant to reject his earlier theodicy of finitude, and that such evil also requires the introduction of a propensity for mendacity to explain it.

Kant's Assertion of a Propensity to Evil

Besides the Collins notes, Kant's fullest discussion of the devilish vices occurs in the 1793-94 Vigilantius lecture notes and the 1797 *The Metaphysics of Morals*. The discussions of the vices in both those places are clearer and more confident than in the Collins notes, but while in both Kant admits the existence of devilish vices (though he does so more clearly in the Vigilantius notes than he does in *The Metaphysics of Morals*), his conception of them as recorded in Vigilantius differs from the one he advances in *The Metaphysics of Morals* (however, I shall not address the discussion in *The Metaphysics of Morals*, as it takes place after the *Religion*, which is all that is relevant to my project in this chapter). What is important about both, though, is that in each Kant posits the existence of a propensity to evil,[49] which is what allows him to admit devilish vices.

To understand how the propensity to evil relates to the devilish vices, it will help to ask the following question: why did Kant, in 1784, claim that the devilish vices did not actually exist but then change his mind about them in 1793? My answer: in 1784 he recognized that some people seemed to exhibit the devilish vices from time to time, but because he denied the existence of a direct inclination to evil (for to admit it would contradict his theory of unsociable sociability), he had to deny that people actually behaved as they appeared to; that is, he had to deny the devilish vices. However, as he recognized in 1791, his denial of a direct inclination to evil (which is roughly equivalent to a propensity to evil), undercut moral responsibility for evil. To render us culpable for

moral evil, he had to introduce a propensity to evil (thereby compromising, to some extent, the efficacy of unsociable sociability as a mechanism for cultural and moral progress). Once he allowed for a propensity to evil, though, he could admit that devilish vice was as it appeared. This is how the introduction of a propensity to evil explains Kant's change of mind on the devilish vices. Let's look right now at some of Kant's advanced thinking on the topic of devilish vices.

In the Vigilantius manuscript, Kant focuses on two innate propensities: a tendency to evil, and a propensity for emulation (note: "tendency to evil" as Kant uses it here is not obviously the same as "tendency to evil" as I defined it in chapter 3; nor is it clearly the same as the "propensity to evil", which encompasses both my sense of the tendency to evil and what I have called the susceptibility to evil; to avoid confusion, then, I shall describe the tendency to evil from the Vigilantius lecture notes as the "V-tendency to evil" or the "V-tendency"). The V-tendency to evil has two components: "a hindrance to {man} in embracing the firm resolve to goodness {and} also … an inner positive ground to evil, {which} constantly arouses in him an inclination towards it" (*LE-V*, 27:571). The V-tendency to evil not only makes it difficult for us to embrace morally good maxims out of respect for the moral law, but also makes us want to adopt evil principles. This makes the V-tendency look quite close to the tendency to evil; the drive to evil, after all, also does two things: it makes certain of our sensible desires judgment-insensitive (i.e., resistant to the diminishing effects of the moral law's judgments); and the effect of this judgment-insensitivity is to embolden the sensible desire that would otherwise have been enervated.

The (natural) propensity for emulation gives rise in each person to a desire to try to equal those around her, either in well-being (happiness) or in non-moral merit (non-moral accomplishments or social standing).[50] This desire to equal others actually stems from the love of honor—the desire for recognition and celebration of one's accomplishments and/or standing—that each of us has. Because we all want to be honored by our fellows, we all strive, both to equal the merit and well-being of others, and to prevent others from eclipsing our own merit and well-being:

> In every man, even in the noblest understanding, there is imprinted a love of honor, and this is the source and principle that finds expression in emulation; so without even thinking of ambition, egotism or selfishness, there nevertheless arises in all – even in well-meaning people – a drive constantly to perfect oneself in comparison with others (*LE-V*, 27:679-80).

The fact that the propensity for emulation is the expression of the innate love of honor, which even good people can have, shows that people acting on the desires this propensity generates are not acting immorally.[51]

Significantly, Kant's discussion of the propensity for emulation gives the impression that it is this propensity that gives rise to the phenomenon of unsociable sociability:

> providence has implanted in mankind an impulse to mutual emulation among themselves in order thereby to compel them to be active in enlarging and cultivating their powers. This easily leads, however, to disparagement in the course of emulation, and thus arises rivalry, or a relation of men to one another that awakens envy of the other's merits, and shuns every danger that might occasion weakness, vis-à-vis the other, want of equal perfection, or even the latter's superiority; we actually hate his very virtues, and feel an inner joy at viewing him with less respect; we notice the other's failings more shrewdly than his merits, in order to put him down. In short, this result of emulation is really a side of human nature that has become malignant, notwithstanding that the purpose of emulation really lay in inciting men to constant cultivation of greater perfection in themselves, by comparison with others. (*LE-V*, 27:678-79)

Providence gives us a propensity for emulation because acting on it allows each of us to enlarge and cultivate her powers. At the same time, it seems as though this propensity for emulation not only causes morally permissible competition (trying to equal or better

others' happiness or merit, or trying to prevent their equaling or bettering your happiness or merit, by improving yourself), but also immoral behaviors. We emulate others in the attempt to gain the plaudits of our peers, but we also disparage them; disparagement gives rise to rivalry, and rivalry in turn "awakens" envy, a devilish vice. Emulation results in "a side of human nature that has become malignant". It looks as though the propensity for emulation is responsible for the devilish vices; indeed, Kant seems to explicitly make this point, claiming:

> *All three of the aforementioned* maxims of viciousness {i.e., the devilish vices} take the ground of their origin from a property of human nature native to man, which not only makes us intrinsically guiltless, but also determines us to an admirable purpose: namely the instinct of antagonism or rivalry, i.e., the inclination to work against the perfections of others, or to surpass them by ever-increasingly promoting our own cultivation, in agreement with the laws of morality. (*LE-V*, 27:692)

Finally, in one place where Kant *does* connect the devilish vices to the V-tendency to evil, he seems to do so only to deny that the devilish vices can come from the V-tendency:

> Devilish vices. Here a man oversteps moral viciousness, or the natural tendency of man towards evil, and thus his tendency to vice is greater than human nature allows, and he seems to have adopted the *principium* of evil itself. Among such vices are ingratitude, envy and *Schadenfreude* {ingratitude with malice}. (*LE-V*, 27:632)

There are two problems with rooting the devilish vices in the propensity for emulation and not in the V-tendency to evil. First, if the devilish vices come from the propensity for emulation, the V-tendency to evil becomes otiose. Second, if the devilish vices stem from the natural, morally blameless propensity for emulation, then Kant's repeated pronouncements that they go beyond (or maybe, below) human nature become impossible to understand.

As for the passage where Kant seems to claim that the devilish vices go beyond the V-tendency to evil, there are two problems with this. First, if even the V-tendency

cannot explain the devilish vices, then it seems Kant must deny that the devilish vices really exist; yet since they seem to exist, such a denial would seem to be *ad hoc*. Second, in some passages, which I have already quoted, Kant appears not only to admit the possibility of devilish vice, he even seems to say they are rooted in the morally innocent propensity for emulation. So Kant seems to be saying both that the devilish vices are rooted in neither the V-tendency nor the propensity to emulation, and that they stem from the propensity to emulation.

Thus, reading Kant's remarks in their most natural fashion forces us to have him contradict himself. If possible, we should strive for a reading that makes sense, not only of all his remarks in the Vigilantius notes, but also of the fact that he seems to have changed his mind on the subject of evil in his theodicy essay. In what follows, I shall offer such a reading.

In my view, the devilish vices are indeed rooted in the V-tendency to evil; consequently, the sense in which they go beyond human nature is intelligible—namely, they go against that part of our nature that providence intended to use to advance the species. Moreover, while it is true that they are *awakened* by the propensity for emulation, they do not spring from it directly. Rather, the love of honor expresses itself through the propensity for emulation, the propensity for emulation activates the V-tendency to evil, and once the V-tendency is activated, people can develop devilish vices.

The first step to arriving at my interpretation begins with examining how Vigilantius's Kant actually defines the three devilish vices. For each of the three vices, he describes an attitudinal version and an aggravated version. The attitudinal version of the vice consists merely in negative attitudes towards others, whereas the aggravated version

pairs the negative attitude with immoral action. "Each of these three inhuman vices is called aggravated, if it is coupled, not merely with aversion for the other's merits and condition, or joy at his misfortune, but also with a desire to damage him in his merits and to contribute actively thereto as author" (*LE-V*, 27:692-93).

Envy, in both its attitudinal and its aggravated versions, contrasts with jealousy[52] (or as Kant also calls it, "misliking" and, somewhat misleadingly, "envy without ill-will" (*LE-V*, 27:694)). A *is jealous* of B when B has more merit than she; because B is more accomplished, or has a higher social status, A feels that her own talents are diminished by comparison. Consequently, she becomes unhappy, and tries to increase her own merit.[53] A *envies* B if she sees B's superior merit as a reason, not to improve her own merit, but instead to hate B. After all, if B weren't there, or were less meritorious than she now is, A wouldn't look worse for the comparison; consequently, although jealousy and envy stem from a similar self-dissatisfaction, jealousy leads to self-improvement (a good outcome), whereas envy leads at least to a negative attitude towards the other, and if it is aggravated, to hostile action.[54]

Whereas envy, in both its attitudinal and aggravated forms, contrasts with jealousy, Kant supplies no word for the innocent precursor to ingratitude. Since Kant eventually (in *The Metaphysics of Morals*) labels this precursor "unappreciativeness" (*MM*, 6:459), I shall do so as well. If B does A a favor (to simplify matters, let us say that A undeniably needs this favor, and does not refuse it when offered), then A, unless she is "angelically virtuous" (*LE-V*, 27:699), will have some feelings of unappreciativeness. This is because B's helping her shows A to be inferior to B in some important way—lacking in skillfulness, wealth, or standing. B's help makes this vivid to A, and as a result

A will feel at least somewhat upset with herself, and will be moved to try to improve her lot.[55] If A is ungrateful, though, then she will hate B for showing himself to be superior to her, and will want to harm B in response.[56]

Someone displays morally innocent *Schadenfreude* if she feels happy at another's misfortune, not because she hates him, but rather because Kant thinks each person's happiness depends to a large extent on how she thinks she is doing relative to others.[57] Thus, if you believe someone around you is doing much better than you, you will feel unhappy because of your relatively paltry well-being. The flipside of this, though, is that if she becomes less happy, your happiness will no longer look so puny, and so you will feel better than you did before. That is, you will take pleasure in her suffering (this is especially true if she was haughty to begin with). *Schadenfreude* becomes exceptionable, though, when someone's superior happiness, and the consequent pain it causes you, makes you want to *cause* her unhappiness.[58] (Note that *Schadenfreude* differs from envy in that, whereas envy is at someone's superior merit, and makes you want to undercut her merit, *Schadenfreude* has to do with someone's superior happiness, and makes you want to diminish her happiness.[59])

As the foregoing shows, there are morally harmless versions of each of the vices that are manifestations of the propensity for emulation. Jealousy, unappreciativeness, and innocent *Schadenfreude* each motivate people to enlarge and cultivate their powers. However, each of these three natural vices can easily transform into envy, ingratitude, and blameworthy *Schadenfreude*, respectively. Though easy, this transformation would be impossible were it not for the V-tendency to evil. In other words, the natural vices (so-

called because they are merely expressions of the natural propensity for emulation) become the devilish vices thanks to the V-tendency.

Several passages support this claim. For instance, Kant writes that the natural vices occasioned by emulation "easily lea[d] … to disparagement in the course of emulation, and thus arises rivalry, or a relation of men to one another that awakens envy of the other's merits" (*LE-V*, 27:678). The propensity for emulation awakens envy, but how does it do that? By giving the V-tendency to evil some materials to work with; without the natural vices, there could be no devilish vices. With the natural vices, though, the devilish vices soon follow. Thus Kant writes that the "result of emulation is really a side of human nature that has become malignant, notwithstanding that the purpose of emulation really lay in inciting men to constant cultivation of greater perfection in themselves, by comparison with others" (*LE-V*, 27:678-79). In other words, the result of the propensity for emulation is an active V-tendency to evil.

How, though, does the V-tendency to evil change the natural vices? To see how the V-tendency works, it will help to have handy the general difference between natural and devilish versions of vices. Someone shows herself to have more merit or well-being than I. In the case of both natural and devilish vice, her being happier or more meritorious than I displeases me. If I am naturally vicious, I try to *improve myself* so that she no longer has more merit or well-being than I; if I am devilishly vicious, I try to *diminish her* so that she no longer has any more merit or happiness than I. It is only this second kind of behavior—diminution of the other—that is immoral.

Now, immorality can come in two flavors: the immorality characteristic of frailty, where a person does something contrary to the moral law while nonetheless judging that

she should not; and the immorality characteristic of depravity, where a person does something contrary to the moral law while also judging that she is doing what she has most reason to do. Call the first kind of immorality "simple immorality" and the second kind "wickedness".

As noted above, the V-tendency makes it difficult for us to embrace morally good maxims out of duty, and it also makes us want to adopt evil principles. Since the V-tendency is what enables us to act devilishly, and since Kant gives no indication that devilish vice is ever done unwillingly (he never describes the envious person as harming the more meritorious *out of weakness*), I take it that the V-tendency to evil makes us want to act wickedly. Given the action-theory of chapter 2, wanting to act wickedly would amount to wanting to accept a motive according to which a sensible incentive should override one's respect for the moral law.

This is strange, though; unless one was already committed to a sensible desire's being more important than respect, why would one want to judge it be *ultima facie* more important? The only explanation I can think of is that the V-tendency to evil operates in the same way as the drive to evil – by strengthening your sensible desires to the point where you wish you had some reason to treat them as more important than your moral obligations. Given that there is no such reason, though, the only way you will be able to persuade yourself of your sensible desire's greater importance is by distracting yourself from the fact that it is not. But how can you conclude that it is permissible even to distract yourself? If you are already evil, you can pull off the trick (as we shall see in chapter 5, this is why having an evil *Gesinnung* must precede acting depraved); but

neither the tendency to evil nor the V-tendency to evil can *make* you side with them. For that, you must already be committed to evil. This is perhaps why Kant writes:

> a tendency to evil is … implanted by nature; but it is clear *ex adductis* that it can also be repressed; so it is certain that, as soon as the tendency is not repressed, but nourished, a degree of imputation arises from this, which is greater than could be occasioned by the tendency previously implanted by nature. (*LE-V*, 27:571-2)

Given this interpretation, we can now explain what Kant means when he says that with the devilish vices "a man oversteps moral viciousness, or the natural tendency of man towards evil, and thus his tendency to vice is greater than human nature allows, and he seems to have adopted the *principium* of evil itself" (*LE-V*, 27:632). The V-tendency to evil only encourages devilishly vicious attitudes and behavior; it is only with the further step, that of nourishing that tendency, that one goes "beyond" the tendency to evil and instead adopts the devilish vices as principles. As Kant later writes:

> In human nature there is a *corruptio*, i.e., all men are susceptible to a {culpable breach of law}. This arises from their *depravity*, a weakness of our sensory nature, and thus is a sin of weakness; but *criminal wickedness*, or the malignity of human nature as such, is acquired, i.e., a deliberate or consciously undertaken transgression of our duties, and this especially according to a maxim we have adopted. (*LE-V*, 27:691)

We should not be mislead by Kant's language; "depravity" here does not mean the same thing it does in the *Religion*; rather, it equates to what Kant in the *Religion* called "frailty"—the susceptibility each person has, whether good or evil, of acting immorally from weakness of will. "Criminal wickedness", on the other hand, *is* depravity in the sense of the term used in the *Religion* (or at least, close to it). What Kant is saying in this passage is that the V-tendency to evil makes everyone susceptible to immoral action, but it is only when the V-tendency is embraced that a person makes one (or more) of the three devilish vices her principle.

So, Kant's story of the devilish vices in the Vigilantius lecture notes is as follows: providence has implanted everyone with a propensity for emulation (i.e., the expression

of each person's love of honor), which moves people to try to better themselves, in the hopes of equaling or surpassing their peers, either in happiness or in merit. However, each person also has a V-tendency to evil, which inclines her to another way to equal or better her peers: namely, through disparagement and aggression. If she accepts the V-tendency's promptings, then she culpably develops in herself devilish vices; without the V-tendency to evil, then, there could be no devilish vice (but of course, the V-tendency also requires a prior depravity).[60]

How does the V-tendency to evil relate to the propensity to evil *qua* tendency? They are almost exactly the same: both are innate; both must be generated through a prior activation for a person to engage in immoral behavior; because both are freely brought upon oneself, both are evil; and finally, both require a prior commitment to evil for their generation. One difference, though, is that Kant *explicitly* recognizes in the *Religion* that the decision to activate one's susceptibility to evil itself seems to require a prior evil. In Vigilantius, by contrast, his position is less clear. Despite this difference, though, both the drive to evil and the V-tendency are required to engage in devilish vice.

7. Kant's Argument for the UT in the *Religion*

I shall take it as established that without a tendency to evil, people could neither be held fully culpable for the immorality they committed (should they commit any) nor could they ever work themselves into perpetrating devilish vice. What remains to be seen is the kind of argument this allows Kant to forward for the UT. Before I get to the argument, though, I need to respond to an objection.

7.1. Does Kant Have a More Direct Route to the UT?

Arguably, Kant has a simple argument available to him to establish the UT, one that has the advantage of needing no historical excavation as a support. Call this the "Simple Argument". The Simple Argument runs like this:[61] (1) the moral law's deliverances appear to all people as commands; (2) the reason these deliverances appear to all people as commands, rather than as actions they want to perform, is that all people have a tendency to evil; (3) therefore, everyone has a tendency to evil.[62] As Kant puts it:

> if we ask, "What is the *aesthetic* constitution, the *temperament* so to speak *of virtue*: is it courageous and hence *joyous*, or weighed down by fear and dejected?" an answer is hardly necessary. The latter slavish frame of mind can never be found without a hidden *hatred* of the law, whereas a heart joyous in the *compliance* with its duty (not just complacency in the *recognition* of it) is the sign of genuineness in virtuous disposition (*Rel*, 6:23n).

What should we say about the Simple Argument? One could argue that (1) is itself a claim that perhaps needs justification; and it is not at all obvious what this justification might be. (One cannot say that all finite beings see the moral law as presenting them with obligations, not just because of the counterexample of angels, but also because Kant allows for the possibility of finite holy beings.)[63] I shall not take this tack, however. I want to focus instead on (2).

The first criticism one may make of (2) is: if it is indeed the case that the tendency to evil is the only thing that can make sense of why we must be obligated to obey the moral law, then why did Kant, in the 1780s, deny the tendency to evil while asserting that the moral law obligates us? Regardless of whether I successfully refute the Simple Argument, its advocates will still have to explain this.

A more direct response to (2) is: I do not think Kant needs to invoke the tendency to evil to explain why it is the moral law presents people with imperatives rather than obligations. Let us return to the action-theory of chapter 2. There, I claimed that everyone

adopts the Moral Maxim and the Prudential Maxim as her two highest disposition-maxims. Because she adopts the Moral Maxim, she judges her duties to be good to do. If she accepted no other maxim, then she would always do her duty; indeed, her "duties" would not appear to her as things she was supposed to do, but as the only thing to do. They would still appear to be good, but it would no longer be the case that they seemed obligatory.

Because she is a finite being, though, she does accept another maxim, the Prudential Maxim. According to this maxim, it is always good to satisfy one's sensible desires. Interestingly, though, if a person accepted only the Prudential Maxim, and not the Moral Maxim, there would still be questions about what she should do, because sometimes sensible desires conflict, and it is not always obvious which of them promises the most pleasure. So, not only would satisfying her sensible desires appear to her to be a good thing, they would also (when she had conflicting sensible desires) appear to her to be (pragmatically) obligatory.

What happens in the real world, where everyone accepts both the Moral and the Prudential Maxim? What happens is that everyone judges both her duties and her sensible desires to be good. Because these two can conflict, though, they also judge them both to be (pragmatically or morally) obligatory. So far, there is no call for the tendency to evil.

The advocate of the simple argument will reply that if a person were *really* good, then whenever a sensible desire conflicted with a moral duty, the sensible desire would lose all attractiveness. If a person is capable of having a sensible desire collide with a duty without being that sensible desire perishing entirely, this must be because something in her spurs the sensible desire on. And since there is no *pro tanto* reason for spurring the

sensible desire on, whatever is in her that invigorates it must be something evil, either the tendency to evil or the evil *Gesinnung* (or both).

But why think this is the case? Why think, that is, that if someone has a sensible desire that conflicts with a duty that the strength of the sensible desire must be completely destroyed? Take, for example, someone who judges her sensible desires to be good, even when they conflict with her moral obligations, but who always judges them to be less good than her moral obligations, and who never acts immorally. Why think that even she has to have something evil in her?

At the point, the advocate of the Simple Argument will claim: sure, she seems good now; but this is only because she has not done anything evil yet. If one simply increases the strength of her sensible desires enough, then at some point she will succumb to them. That is, the strength of her sensible desires must eventually reach a point where they persuade her to change her judgment of their goodness from "good, but not as good as following the moral law" to "good, and better than obeying the moral law". After all, this is how the tendency to evil works: by strengthening certain sensible desires, it increases the likelihood that a person will succumb to them. If you, the advocate of the non-simple argument for the UT, were to deny this, then you would also have to revise your understanding of the tendency to evil.

I (an advocate of the non-simple argument) do not think this is the case, though. True, by strengthening a person's sensible desires the tendency to evil inclines her to draw the judgment that they do indeed provide reasons that override those provided by the moral law. But she actually draws this judgment only if she has an evil *Gesinnung*. A tendency to evil without an evil *Gesinnung* would never move a person to conclude that

her sensible desires sometimes provide her with practical reasons that supersede those given by the moral law – she would merely have particularly strong desires (she could even occasionally succumb to those desires, though this would require momentary self-deception).

Where does this leave us with regard to (2)? We have no good reason to believe (2). It seems possible for someone to experience the deliverances of the moral law as binding, even if she has no tendency to evil, while nonetheless being a good person (i.e., being such that she will never succumb to her sensible desires or judge them to have greater normative authority than her moral obligations). All it takes for a person to qualify as good is being such that she never judges her sensible desires to present her with better reasons for action than her moral obligations. Consequently, the bare fact that people experience categorical imperatives is not enough reason to think that they must have a tendency to evil.

7.2. Kant's Argument(s) for the UT

The Simple Argument for the UT does not work. What, then, are Kant's arguments for the UT? I think there are two: what I call "the Wholeheartedness Argument", according to which we can reason from the fact that all people can (on some occasions at least) wholeheartedly engage in immorality to the conclusion that everyone has a tendency to evil, and "the Argument from Devilish Vice", which begins with the premise that everyone is capable of devilish vice and moves to the conclusion that, therefore, everyone has a drive to evil. Let us examine these arguments more closely.

7.2.1. The Wholeheartedness Argument

Kant's original theodicy of finitude exonerates God by seeing all immorality as stemming from frailty. If it were possible for us to be morally responsible for something more – for full-blooded evil-doing, i.e., depravity – then we would have to have a germ to evil. But God could have no good reason for giving us a germ to evil (for one thing, stocking us with such a germ would contradict God's providential plan, which unfolds thanks to unsociable sociability); consequently, God did not give us any such germ, so we have no tendency to evil.

If one assumes that people *are* capable of full-blooded immorality – if, that is, one holds that you and I and everyone else are capable of judging our moral obligations to have less normative authority than certain of our sensible desires and acting on that judgment – then one must hold that everyone has a tendency to evil.

But it seems false that everyone is capable of full-blooded immorality. After all, some (precious few) of us have good *Gesinnungen*; people with good *Gesinnungen* do not *ever* judge any of their moral obligations to be less worthy of fulfillment than any of their sensible desires – if one did, then, *ex hypothesi*, she would not be good, because to make such a judgment one has to have an evil *Gesinnung*. Good people, then, do not seem to be capable of judging their moral obligations to be less authoritative than their sensible desires.

We must analyze what it means to be capable of something, though. While it is true that no individual, *when she is good*, could ever engage in full-blooded immorality, it is possible for any good individual to relapse into evil.[64] This is because of the tendency to evil; even a good person, because of her propensity to evil *qua* tendency, can retread the steps of the path by which she originally became depraved.[65] In this sense, then, all

people with a tendency to evil are capable of engaging in full-blooded immorality (and being culpable for it when they do).

7.2.2. The Argument from Devilish Vice

It seems that people engage in devilish vice. That is, it seems that people sometimes act immorally, not out of weakness of will, but because they want to improve their comparative position. Devilish vice is just a species of depraved action, a kind of action where someone engages in full-blooded evil. As was just shown in section 7.2.1, though, for everyone, even good people, to be capable of depraved action, everyone must have a tendency to evil, something that can tempt her to move her back from good to evil, or that can further her along the path from frailty to impurity to depravity.[66]

Let us more put the Wholeheartedness Argument and Argument from Devilish Vice more formally. They are:

Wholeheartedness Argument

1. Everyone is capable of wholehearted immorality.
2. In order for a person to be capable of wholehearted immorality, she has to have a tendency to evil.
3. Therefore, everyone has a tendency to evil.

Argument from Devilish Vice

4. Everyone is capable of devilish vice.
5. In order for a person to be capable of devilish vice, she has to have a tendency to evil.
3. Therefore, everyone has a tendency to evil.

What a formal presentation of these two arguments shows is that they are more or less the same argument. The only difference between the two is the replacement of "devilish vice" with "wholehearted immorality". Thus, I think that the problem Kant had with the germ to evil in his lectures on religion was the same problem he had with the direct inclination to evil in his lectures on ethics. In both cases, Kant found it difficult to admit the possibility of knowing violation of the moral law—either because it defeated his theodicy or because it conflicted with his theory of historical development (and let us not forget that Kant intends his theory of historical development as a theodicy; see *IUH*, 8:30).

One should ask how we know the truth of 1 or 4. I answer this question in the next section.

7.3. "We Can Spare Ourselves the Formal Proof …"

We can now get to the final part of my argument, which is my speculation about why Kant spared us the formal proof he seemed to have up his sleeve. If he had it, why not present it?

I conjecture that Kant spared us the proof both because of the nature of his proof and because the nature of his audience. First, let us look at the proof itself.

Both the Wholeheartedness Argument and the Argument from Devilish Vice are valid, deductive arguments. Yet they both suffer from a crucial flaw: the first premise of each is, strictly speaking, false. This is because there is a state of affairs, namely that exemplifying the highest good, in which it is false that everyone is capable of either wholehearted immorality or devilish vice.

Although the first premises of both arguments are false when "everybody" is construed strictly, to include all human persons, they lose their appearance as transcendental arguments when rephrased correctly. Correctly rephrased they run:

Wholeheartedness Argument*

1* Everyone in an evil society is capable of wholehearted immorality.

2* In order for a person to be capable of wholehearted immorality, she has to have a tendency to evil.

3* Therefore, everyone in an evil society has a tendency to evil.

Argument from Devilish Vice*

4* Everyone in an evil society is capable of devilish vice.

5* In order for a person to be capable of devilish vice, she has to have a tendency to evil.

3* Therefore, everyone in an evil society has a tendency to evil.

As these re-phrasings show, these are not transcendental arguments – because of their first premises, they do not have the mark of universality. In other words, one reason Kant could spare us the formal proof is that he did not have one.

But he had something near enough; for surely it is hard to doubt that all the people we see around us, even if they do not actually engage in wickedness, are *capable* of engaging in it. Why might someone doubt this? In theory, she could advert to her hard-headedness – Kant does not apodictically prove that every human person we will ever experience is evil, so his argument fails – but is this really what is going on? After all, if someone had an argument that relied on a premise as, well, *obvious* as 1* or 4*, and used

it to prove a profound truth about all of us, would you really worry too much about that premise?

You might, if the truth of this premise showed you to have to do something onerous. And of course, if 1* or 4* is true, you *do* have to do something onerous: you have to take up arms against yourself; you have to ceaselessly struggle for moral self-improvement and descend "into the hell of self-cognition" (*MM*, 6:441).

The problem is, to sincerely deny 1* or 4* you would have to be completely innocent or arguing in bad faith. For such people, argument will not work; to convince them, you must present them with observable evidence for the ubiquity of the tendency to evil. This is precisely what Kant does; and moreover, as I shall show in the next chapter, he thinks most of us are indeed such people.

---

[1] See A85/B117.
[2] See section 3.3.1 of chapter 3.
[3] See section 3.2 of chapter 3.
[4] See section 3 of chapter 3.
[5] See section 4.2 of chapter 3.
[6] To be fair, Allison writes that Kant's claiming that we can spare ourselves the formal proof is "a rhetorical ploy designed to enable Kant to bracket the deeper questions posed by the problem" of proving the universality of the propensity to evil (Allison 2002, 341). However, since Allison never explains why Kant at this point needs to bracket the deepest questions, I do not take him to have made much headway on this issue.
[7] The passage I have in mind runs:

> is man by nature morally good or bad? He is neither, for he is not by nature a moral being. He only becomes a moral being when his reason has developed ideas of duty and law. One may say, however, that he has a natural inclination to every vice, for he has inclinations and instinct which would urge him one way, while his reason would drive him in another. He can only become morally good by means of virtue—that is to say, by self-restraint—though he may be innocent as long as his vicious inclinations lie dormant. (*E*, §102)

[8] Most of the notes of these lectures were taken from lectures given between 1783 and 1784, but parts of them (most likely 28:1077-80, 1088, 1100, 1113, and 1116) were probably from lectures given between 1785 and 1786 (see *Religion and Rational Theology*, 337-38).
[9] See *Lectures on Ethics*, xv.
[10] Similarly, Kant asks, "Shall we derive evil from a holy God?" (*LPDR*, 28:1077) and answers, "who but the human being is responsible for {evil}? This way of understanding things agrees completely with the mosaic story" (*LPDR*, 28:1077).
[11] See also Kant's remark that "The man who deems himself unfortunate is also malicious, for he envies others their good fortune … but he who in misfortune still shows a cheerful and steadfast spirit, who maintains a firm courage, even when he has lost everything, still has that within him which possesses an intrinsic worth, and such a man deserves compassion instead" (*LE-C*, 27:367-68).

[12] For an excellent statement of a free will theodicy, see van Inwagen 2006. It should be noted, though, that van Inwagen sees what he does as a "defense" rather than a "theodicy".

[13] Kant's emphasis on freedom as the condition of all other goods leads to a peculiar problem for him, one he never addresses: given that God has freedom, and given that he creates the world, why is the world, the free product of a perfectly good being, not intrinsically good? In other words, why do there need to be people besides God in the world for it to have value?

As a first pass at answering this question, I bring up the following considerations. First, just because something is the product of a free, perfectly good being, it does not follow that *every aspect of it* is good. For instance, it would be good for a good person to freely give alms to the poor, but it does not follow that her giving the alms with her left hand was good; which hand she uses to give the alms is of no moral value whatsoever. Similarly, even though God's creating the world is a good act, it does not follow that everything in the world God creates has moral value. To find out what aspects of God's creative act have moral value, then, we should ask why God would want to create the world at all. Kant's answer is the Leibnizian one that God creates, not to bring about just any world, but the best of all possible worlds, i.e., the highest good. In order to do that, though, God needs to create free, finite creatures (free, so they can guide themselves according to the moral law; finite, because it is a logical contradiction for God to create another infinite being). Thus, in order to create a world with moral value, God needs to create human beings.

One could of course ask why God should create a world at all. One answer that Kant does not give, but that seems to comport with his approach to these issues, is found in Adams 1987. According to Adams, God cannot have the virtue of grace unless there are imperfect creatures. Thus, God creates in order to have grace; this dovetails with Kant's emphasis on God's grace as the instrument by which people can effect moral revolution in themselves (*Rel*, 6:44, 118), which revolutions are necessary conditions for the highest good coming about.

[14] Cf. *LPDR*, 28:1113.

[15] It would be impossible for God to create another infinite being, for if that being were truly infinite, it would have all the same characteristics of God, one of which is necessary existence, and a necessary being cannot be created, for that would imply that it was in some sense dependent on something else.

[16] Allison calls this "originary freedom" (Allison 2002, 343). Derk Pereboom expresses the intuition underlying Kant's position (which he, not referring to Kant, calls "source incompatibilism") with the following conditional: "If an agent is morally responsible for her deciding to perform an action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which the she has no control" (Pereboom 2006a, 2).

[17] This interpolation comes from Pereboom 2006b, 559.

[18] Moreover, he asserts the existence of diabolical vice in his 1797 *The Metaphysics of Morals*; see especially *MM*, 6:458-61.

[19] The attentive reader may have noticed that Kant describes the devilish vices as "simply the idea of a maximum of evil that surpasses humanity". This may suggest to her that Kant does not, in fact, really think that the diabolical vices exist. However, I show in section 5.2 of this chapter that "surpassing humanity" does not mean "impossible for humans", but rather "conflicting with nature's designs for humanity".

[20] See *LE-C*, 27:380 and 439.

[21] As John Deigh writes, according to the conventional analysis of things (from which he dissents), "shame goes to failure, guilt to transgression. Shame is felt over shortcomings, guilt over wrongdoings" (Deigh 1983, 225).

[22] The examples given might leave the impression that it is only when someone is *unhappy* and others are positively *happy* that one begrudges others their happiness. After all, I begrudge others their good spirits when I am *dis*contented; I begrudge others their *good* fare when I have *poor* fare; and I resent everyone else's living when I alone have to die. On this view, you have to be unhappy in the first place to hold a grudge. Kant seems to support this view, writing, "when a good-natured person is happy and cheerful, he wishes that everyone in the world might be equally happy and cheerful, and begrudges it to nobody" (*LE-C*, 27:439). Notice, though, that even the happy person wants others to be only *equally* happy and cheerful, not happier.

[23] Just because there is a grudging *element* in our nature, it does not follow that grudging can never be morally culpable. "Grudging is more natural {than envy}, though it, too, should not be condoned" (*LE-C*, 27:438).
[24] Definition found at http://dictionary.oed.com.proxy.lib.umich.edu/cgi/entry/50076520?query_type=word&queryword=envy&first=1&max_to_show=10&sort_type=alpha&result_place=1&search_id=x8WD-acD1l5-12502&hilite=50076520.
[25] See also *LE-C*, 27:443: "Envy does not consist in wanting to be the happiest, as with grudging, but in wanting to be the only one happy. This is the vilest thing about it, for why should not others be happy too, when I am?"
[26] See *LE-V*, 27:695.
[27] However, in the Collins lecture notes Kant also says that *Schadenfreude* is "a sort of animality, whereby man retains something of the beast in him, which he cannot overcome" (*LE-C*, 27:441). If *Schadenfreude* is merely an animal tendency that cannot be overcome, then it is not clear that it is a vice at all. If that is the case, then devilish *Schadenfreude* would be an enjoyment of others' suffering that one could refrain from, while natural *Schadenfreude* would be an uncontrollable instinct.
[28] I assume that this passage has to do with *Schadenfreude* rather than envy; if he were talking about envy, I should think he would mention the envious person's desire to get all happiness for herself. This is supported by the fact that Kant later (*LE-C*, 27:443-44) explains away diabolical *Schadenfreude* the same way he does here.
[29] Assuming there is such a thing as diabolical *Schadenfreude* in the first place. If there is not, then one would have to interpret Kant as finding normal *Schadenfreude* as expressive of a direct inclination to evil, which he then reinterprets to express only an indirect inclination to evil.
[30] Kant does not attempt to reinterpret devilish envy, but he could tell the following story: The diabolically envious person seeks to ensure that she is the only happy person, not because she thinks of herself as superior, but rather because she believes that other people who are happier than she is think of themselves as superior to her. Thus, if she does not do what she can to deprive them of their happiness, they will take the opportunity to try to reassert their superiority over her.
[31] For a detailed explanation of how the propensity to evil enables this, see chapter 3 of this dissertation.
[32] It is anachronistic to project the predispositions to humanity and personality, which Kant articulated only in 1793, back into this 1784 essay; however, as the First Proposition indicates, Kant did have a notion of predispositions, and although the predispositions he had in mind were probably the "technical", "pragmatic", and "moral" predispositions (see *Ant*, 7:322-24), they are close enough to the predispositions to humanity and personality to render the anachronism harmless.
[33] See *LE-C*, 27:390-91.
[34] See *LE-C*, 27:391-92.
[35] See *OM*, 8:258-63.
[36] See *OM*, 8:265-67.
[37] See *OM*, 8:257.
[38] "The natural motivating forces for this, the sources of unsociability and continual resistance from which so many ills arise, but which also drive one to the renewed exertion of one's energies, and hence to the further development of the natural predispositions, thus reveal the plan of a wise creator" (*IUH*, 8:21-22).
[39] See *OM*, 8:259.
[40] This shows that, while source-incompatibilism is perhaps a route Kant could have taken to allow for evil coming both from our finitude and from our spontaneity, it is either not the route he actually took, or one he considered but eventually found wanting.
[41] To be fair, Kant does not think in his lectures on religion that our desires *literally* overpower us; after all, as was shown in chapter 2, section 2.2.3.2, stimuli only affect us, they do not determine us. What really must happen is that we let our desires compel us. *Why* we let them do this, without our having a tendency to evil, though, is apparently not something that Kant ever explored before the *Religion*.
[42] It might also explain his notorious remark in the *Groundwork* that "the inclinations themselves, as sources of needs, are so far from having an absolute worth, so as to make one wish to have them, that it must instead be the universal wish of every rational being to be altogether free from them" (*G*, 4:428), as well as his claim in the second *Critique* that "the inclinations change, grow with the indulgence one allows them, and always leave behind a still greater void than one had thought to fill. Hence they are always

burdensome to a rational being, and though he cannot lay them aside, they wrest from him the wish to be rid of them" (*CPrR*, 5:118).

[43] Or so Kant seems to argue (see *OM*, 8:265-68); it is not clear, though, why Kant is entitled to this claim (or even to the lesser claim that most people who claim to know God's reasons are not being sincere). After all, people might not know of Kant's "proof" of the inaccessibility of God's reasons; they might have been raised within a religious tradition that stressed a particular theodicy, and they might furthermore never confront someone who challenges it. Surely such a person could sincerely, even if wrongly, claim to know God's reasons for permitting evil.

[44] See *OM*, 8:257.

[45] Indeed, Kant's nine different theodicies call to mind his remark in the first *Critique* that metaphysics is a "battlefield of … endless controversies" (A viii).

[46] "'{M}aterial conscientiousness' consists in the caution of not venturing anything on the danger that it might be wrong, whereas 'formal' conscientiousness consists in the consciousness of having applied this caution in a given case" (*OM*, 8:268).

[47] This is reasserted in *MM*, 6:429-31.

[48] In this, it is like the propensity for evil *qua* tendency. See chapter 3, section 3.3.3.

[49] In *The Metaphysics of Morals*, Kant posits the propensity to evil at *MM*, 6:380n. See also *Anthropology from a Pragmatic Point of View*, where Kant asserts a "tendency to evil" at *Ant*, 7:324.

[50] See *LE-V*, 27:678.

[51] Admittedly, it is a reversal of things to see a propensity as the expression of an inclination; one way of making sense of this is by interpreting "propensity" not in Kant's technical sense, but as a mere tendency to behave.

[52] See *LE-V*, 27:693.

[53] "If the other's advantages arouse merely distress in a man, because on comparing his worth with the moral standing of the other he feels himself degraded, this is merely *misliking* or {envy without ill-will}. He feels merely his own unworthiness by the comparison made" (*LE-V*, 27:694).

[54] "{E}nvy becomes {malicious envy}, i.e., {spite}, when within him there is simultaneously awakened the desire to lessen those advantages, and to injure the other on that account" (*LE-V*, 27:694).

[55] "Ingratitude *in genere* (*ingratitudo*) is likewise a displeasure or discontent at the obligation which the other has laid on us, through the kindness he has shown towards us" (*LE-V*, 27:694-95).

[56] Ingratitude "becomes {malicious} if from {kindness} there arises a hatred for the well-doer, and a passion for doing him harm and evil, just because he has conferred benefits upon us" (*LE-V*, 27:695).

[57] See *LE-C*, 27:366-68.

[58] *Schadenfruede* "is {malicious} if it is coupled with a desire to render the state of the other unhappy" (*LE-V*, 27:695).

[59] Attitudinal or aggravated *Schadenfreude* "differs … from envy of the same type, in that it seeks to lower, not the worth of a person, but his state of happiness" (*LE-V*, 27:695).

[60] This raises the question: if the activation of the V-tendency (or for that matter, the drive to evil) requires a prior commitment to evil, why does Kant need to bring in the V-tendency (or drive) to evil in the first place? I shall explain this in section 4 of chapter 5.

[61] I owe this argument to Brian Chance.

[62] This is close to the argument Allison gives in Allison 2002, 342.

[63] See *MM*, 6:383.

[64] See section 4.1 of chapter 3.

[65] As will be shown in chapter 5, I take there to be two "moments" of depravity: one's initial depraved act of donning an evil *Gesinnung*, and the later moment of manifesting a particular kind of depravity that happens only after many instances of frailty and impurity.

[66] See *Rel*, 6:41-42.

Chapter 5: How to Be Evil: Kant's Moral Psychology of Immorality

1. Introduction

As was shown in chapter 4, by the 1790s Kant became convinced that people sometimes engage in actions they know to be wrong. I call such actions "wholeheartedly immoral", or "wicked" actions. To explain how people can engage in such actions, Kant posits a tendency to evil. What I have not yet shown, though, is *how* this drive to evil allows people to act in ways they know to be wrong. Showing this is the main aim of this chapter.

But I have another aim as well, and this is to respond to Claudia Card's criticism that Kant "does not acknowledge immoral principles other than that of subordinating morality to prudence" (Card 2002, 84). If Card is right, then Kant's theory of evil suffers from a significant weakness, because there do appear to be principles that are immoral, but not because they subordinate morality to happiness. For example, "Nationalist principles may be neither prudential nor moral but capable of conflicting with both" (Card 2002, 84). In other words, Kant's theory of evil is too simple to account for the variety of evils we see.

2. The Problem of Willful Evil

It might seem that on Kant's view there can be no such thing as willful evil. This is because people concede the paramount authority of the moral law:

> I do not … need any penetrating acuteness to see what I have to do in order that my volition be morally good. … Although I do not yet *see* what this respect is based upon (this the philosopher may investigate), I at least understand this much: that it is an estimation of a worth that far outweighs any worth of what is recommended by

> inclination, and that the necessity of my action from *pure* respect for the practical law is what constitutes duty, to which every other motive must give way because it is the condition of a will good *in itself*, the worth of which surpasses all else. (*G*, 4:403)

Not only do all of us, including the most evil,[1] accede to the overridingness of the moral law's ukases, it seems that none of us can forget them:

> Every human being has a conscience and finds himself observed, threatened, and, in general, kept in awe (respect coupled with fear) by an internal judge … It follows him like his shadow when he plans to escape. He can indeed stun himself or put himself to sleep by pleasures and distractions, but he cannot help coming to himself or waking up from time to time; and when he does, he hears at once its fearful voice. (*MM*, 6:438)

Everyone concedes that she always has most reason to do whatever her duty is. Moreover, whenever anyone is deliberating about what to do, her conscience arises to warn her off immoral actions and motion her to her obligations by reminding her of her commitment to the moral law's nonpareil authority.[2] It seems that the only way a person can undertake an immoral action is by forgetting its immorality. If this is the case, then no one can wholeheartedly undertakes an action she knows to be immoral. But if no one can engage in wickedness, then my analysis in chapters 3 and 4 of depraved actions as involving the judgment that one has more reason to satisfy one's sensible desire than one does to do one's duty, must fall by the wayside.

I agree that it is impossible to reason as follows: "*A* is morally obligatory, and I have overriding reason to do what is morally obligatory; nonetheless, I think it is better to do *B*." Obviously, if you really think *A* is supported by a reason that overrides all others, then you cannot simultaneously think that it may be overridden by some other reason. However, there are three other ways to judge a sensible desire to be more worth satisfying than a moral obligation.

First, you can act wrongly from ignorance. Suppose some action *A* is, as a matter of fact, morally obligatory but you do not know this. If this were your situation, you

could see some other action *B*, an action on a sensible desire, as better to do. This, however, would not be willful evil.[3]

Second, you can act immorally out of allegiance to a false law. Suppose *A* is really morally obligatory, i.e., obligatory according to the moral law, but you happen to think the law according to which it is morally obligatory (i.e., the moral law) is a false law. You might think this either because you think some other law is the real moral law, or because you do not believe in morality at all. If this were the case, you could reason, "I know that *A* is 'morally obligatory', but I nonetheless think it is better to do *B*."

Third, you can act immorally because of a false belief about your status in relation to the moral law. You could think, "I know the moral law is the true law, and I know that *A* is obligatory according to the moral law, but I am exempt from the moral law as far as *A* goes, because I am special in such-and-such a way. So, while everyone else has overriding reason to do *A*, for anyone who is special, like I am, it is better to do *B*."

Kant thinks that anyone who fails to abide by her obligations, either because she is allied to a false law or has a false impression about her moral status in relation to the real moral law, acts depravedly. Now, one could challenge this view. For instance, it could be that a person adheres to a false law or has a false belief about her moral status, but does this inculpably. Kant, though, thinks it is impossible not to have known about the real moral law and its paramount authority *at some point*: "In regard to his natural obligations, nobody can be in error; for the natural moral laws cannot be unknown to anyone, in that they lie in reason for all" (*LE-C*, 27:355). So the reason Kant thinks that a person who adheres to a false law or has a false belief about her moral status is depraved is that anyone who does either of these things (i.e., adhering to a false law or having a

false belief about her moral status) is morally culpable for putting herself into that condition. In the rest of this chapter, I explore in detail how a person can put herself into this condition and why anyone who does this should be seen as having an evil *Gesinnung* and as capable of wholehearted evildoing.

3. Some Preliminary Distinctions

In this section I work through a thicket of interrelated concepts, explaining what they are and how they relate to each other. The concepts I have in mind are: inner worth, moral worth, and relative worth; love and respect; self-love and self-respect; and pragmatic merit and moral merit. In order to see how a person becomes depraved, one must understand these concepts.

Worth

Kant distinguishes among three kinds of worth: "inner worth", "moral worth", and "relative worth" (it should be noted that he sometimes uses inner and moral worth synonymously; in general, though, the main thrust of his usage is as I describe it in this section). Inner worth is the worth a person has as a result of having the capacity to set herself ends and act from respect for the moral law[4] (another name for inner worth is "dignity")[5]. Although Kant sometimes calls this capacity "personality",[6] he also refers to it as "humanity"[7] (following standard convention, I shall also call it humanity). Because every rational agent has humanity, every agent has inner worth; moreover, because humanity is not a capacity that comes in degrees but is instead something that one either does or does not have, every agent has equal inner worth.[8]

Moral worth is the worth a person has as a result of how she uses her humanity: "Personal self-assessment, or the determination of one's own moral worth, the {just

estimate of oneself}, rests on a comparison of one's action with the law" (*LE-V*, 27:703).[9] As this passage shows, people measure their moral worth by comparing their conduct to the moral law. The more closely aligned their conduct with the moral law, the more moral worth they have. Thus, people have different degrees of moral worth; those who have evil *Gesinnungen* have no moral worth,[10] while those who have good *Gesinnungen* have more or less moral worth depending on how able they are to withstand frailty and impurity.[11]

Inner worth is a property only of people, but both people *and actions* can have moral worth.[12] However, while people have varying degrees of moral worth, actions either do or do not have moral worth (i.e., they either are performed out of respect or they are not) – they do not vary in the degree of their moral worth[13] (though as we shall see in section 3.4 of this chapter, they can differ in their degree of merit).

Finally, relative worth is, like moral worth, a comparative measure, but whereas a person's moral worth is determined by how she compares to the moral law, her relative worth depends on the interests of those who are trying to evaluate it.[14] Thus, if I am interested in advancing my academic career, and meet a mechanic, she is worthless for me relative to that end; but if I am interested in fixing my car, then she is comparatively worthier than most other people. (Note that both people and actions can have relative worth, and that the relative worth of both comes in degrees, depending on how much either advances the interest of the person determining the relative worth.)

So far I have discussed the kinds of worth under what you might call "ideal conditions" – that is, as pertaining to what they are supposed to pertain to. Only humanity has inner worth, only morally good conduct has moral worth, and only the useful has

relative worth. Although this is how things are supposed to be, we can imagine things being different. For instance, a person could understand what inner worth is, but nonetheless think that some people have more of it than others; for instance, someone could think that the humanity of the Brahmin has more inner worth than that of the untouchable, or that moral worth should be determined, not by comparison to the moral law, but by comparison of one person to another.[15] This will be important for understanding both the kinds of false laws people can believe instead of the moral law, and the kinds of beliefs they can have about their own status relative to the moral law.

Respect and Love

The different kinds of worth call for different kinds of attitudes. The proper reaction to something on account of its inner or moral worth is respect or esteem, while the proper attitude to something because of its relative worth is love:

> We esteem what has an inner worth, and love what has worth in a relative sense; understanding,[16] for example, has an inner worth, regardless of what it is applied to. The man who observes his duty, who does not degrade his person, is worthy of esteem; the man who is companionable is worthy of love. (*LE-C*, 27:357-58)

Recognition and Appraisal Respect

What is it to respect or esteem something, and why do inner and moral worth deserve it? As it turns out, Kant seems to understand respect in different ways, depending on whether it is a response to inner or to moral worth. The kind of respect we are supposed to have for someone on the basis of her dignity manifests as a kind of behavior, specifically, the behavior of always treating her as an end in herself and never as a mere means.[17] Following Stephen Darwall, I shall call this kind of respect – the kind of respect that "consists in giving appropriate consideration or recognition to some feature of its object in deliberating about what to do" (Darwall 1977, 38) – "recognition respect".

On the other hand, the kind of respect we are supposed to have for someone on the basis of her moral worth is not something that necessarily must manifest as a particular kind of behavior, but is rather an attitude of appraisal. About this kind of respect Kant writes, "We acquire respect {for someone} in virtue of good conduct" (*LE-C*, 27:409-10). In other words, this variety of respect is supposed to track a person's moral worth: the more moral worth someone has (or displays in an action),[18] the more of this type of respect we are supposed to have for her. Taking Darwall's cue again, I shall call this "appraisal respect" (Darwall 1977, 39).

I wrote that recognition respect *manifests* as a kind of behavior, but I have not yet described the attitude that it manifests. Kant writes of the attitude:

> from our capacity for internal lawgiving … there comes *exaltation* of the highest self-esteem, the feeling of his inner worth (*valor*), in terms of which he is above any price (*pretium*) and possesses an inalienable dignity (*dignitas interna*), which instills in him respect for himself (*reverentia*). (*MM*, 6:436)

As this passage makes clear, recognition respect is the same attitude I described in chapter 2 as plain, old "respect". Thus, recognition respect is the feeling of awe at something's greatness, which feeling makes a person want to treat the locus of greatness in a very deferential way. Feeling recognition respect for something does not, of course, guarantee that you will treat it as an end in itself; however, if you recognition-respect something but treat it as a mere means to one of your ends, then you will feel a deep sense of self-contempt for having subordinated something of such great importance to something of such comparatively little worth. In Kant's words, "Contempt … is unbearable. An object of contempt is despised by everyone. It takes away our worth for others, and also the consciousness of our own worth" (*LE-C*, 27:407).

Such exaltation is supposed to be the attitude that gives rise to recognition respect. However, people can recognition-respect others from a different motivation, namely that of obedience to the state or fear of punishment. Such a person would happen to treat others as ends in themselves, but not out of awe. Kant writes of such respect that:

> It is not to be understood as the mere *feeling* that comes from comparing our own *worth* with another's … It is rather to be understood as the *maxim* of limiting our self-esteem by the dignity of humanity in another person, and so as respect in the practical sense (*MM*, 6:449).

Besides describing its object and how it co-varies with it, I have not said much about appraisal respect. It cannot be phenomenologically quite the same as recognition respect, for recognition respect is directed to humanity, the source of all value and literally the most important thing in the universe.[19] Whereas recognition respect is awe coupled with deference, appraisal respect must be more like a feeling of approval, or admiration, for it is directed to people who have done things to promote humanity, rather than to humanity itself.

One must note, however, that there are different reasons we can have for approving of or admiring someone. Suppose someone is gifted in some way and makes good use of that gift. I can approve of her making good use of her gift because it will benefit me; this would be a self-interested kind of approval. Alternatively, I can approve of her putting her talent to good use, not because it will help me, but rather because that is what one is supposed to do with one's talent; this would be a selfless approval. Appraisal-respecting someone feels like this selfless kind of approval, with one significant addition. Since everyone is capable of living up to her humanity and is furthermore supposed to do it,[20] whenever someone does the right thing, an observer cannot help but to use that demonstration of moral worth as an occasion to think of her own. If you have a high

sense of your own moral worth, then you will feel solidarity when you see someone else demonstrate her own. If you do not have a sense of your moral worth (say, you are readying yourself to face a situation where you will have to morally exert yourself), you will view an instance of morally worth action as motivating for yourself – after all, if she can do it, then so can you. Finally, if you have a low sense of your own moral worth, then others' displays of it will humiliate you.[21]

Love of Well-Liking and Love of Well-Wishing

Love is an attitude properly directed to things of relative worth. "We … love what has worth in a relative sense … the man who is companionable is worthy of love" (*LE-C*, 27:357-58). If I find someone companionable, I will love him, because his company gives me pleasure. My love, however, is contingent upon his having a worth for me; should I cease finding his companionship enjoyable, then, because he will no longer have any worth for me, I will stop loving him.

Love does not have a single meaning, though. As with respect, there are two kinds of love – an attitudinal form that Kant calls "love of well-liking" (*Wohlgefallen*) and a behavioral form he calls "love of well-wishing" (*Wohlwollen*). "All love is either love that wishes well, or love that likes well. Well-wishing love consists in the wish and inclination to promote the happiness of others. The love that likes well is the pleasure we take in showing approval of another's perfections" (*LE-C*, 27:417).

We have well-liking love for people or for their perfections when we take some kind of positive attitude (a "liking") to them because they or their perfections have a worth for us.[22] Well liking:

> may be either sensuous or intellectual. All such liking, if it is love, must first of all be inclination. The love that is sensuous liking is a delight in the sensuous intuition, due to

> sensuous inclination; sexual inclination is an example of this; it is directed, not so much to happiness, as to the mutual relation of the persons. (*LE-C*, 27:418)

I have sexual liking for someone if, because of her sexually attractive properties, I desire sexual congress with her. Notice, though, that if this woman were, from the perspective of my sexual desire, relatively worthless for me, I would not have well-liking love for her.

I show well-wishing love for someone if I act to benefit her. (In *The Metaphysics of Morals* Kant calls well-wishing love "beneficence" (*MM*, 6:402).) I can wish someone well for all sorts of reasons – because she has benefited me, because it gives me pleasure to benefit her, because someone has paid me to do so, etc. – though importantly, Kant thinks that being beneficent to someone for long enough will result in one's holding a particular attitude to her, namely love of well-liking or "benevolence" (*MM*, 6:401): "If someone practices {beneficence} often and succeeds in realizing his beneficent intention, he eventually comes actually to love the person he has helped" (*MM*, 6:402).

Importantly, benevolence or love of well-liking is more general than the attitude we nowadays call love. Rather than encompassing specifically romantic, familiar, or collegial feelings it seems that it can refer to any positive feeling for a person or her properties (besides, of course, respect). It seems that love of well-liking is directed either to perfections (i.e., good properties) or to a person on the basis of her perfections:

> There is … a distinction to be drawn in a man between the man himself and his humanity. I may thus have a liking for the humanity, though none for the man. I can even have such liking for the villain, if I separate the villain and his humanity from one another; for even in the worst of villains there is still a kernel of good-will. (*LE-C*, 27:418)

We should note again that the above describes "ideal conditions" for respect and love. There does not seem to be anything impossible about having recognition or appraisal respect (or at least something analogous) for something that does not deserve

it,[23] and one could fail to show recognition or appraisal respect for something that does. Similarly, one could have a love of well-liking toward that for which one should have appraisal or recognition respect, and appraisal or recognition respect toward that which one should merely love.

Self-Love and Self-Respect

So far, I have been discussing love and respect as addressed to others. But Kant says a lot about love and respect as directed to oneself, i.e., self-love and self-respect. Let us start first of all with self-respect.

Recognition Self-Respect and Appraisal Self-Respect

Just as the attitude of recognition respect mandates that you treat people in certain ways – namely, always as ends in themselves – so too self-directed recognition respect ("recognition self-respect") requires you to treat yourself in certain ways. For instance, because you have dignity, you have a perfect duty to refrain from killing yourself for the mere purpose of averting pain.[24] Furthermore, and (once again) as with other-directed recognition respect, the reason recognition self-respect suggests treating yourself as an end in yourself is that your humanity has inner worth. Consequently, you view your own humanity with a mixture of awe and deference, just like you would view anyone else's.

One important difference between recognition self-respect and other-directed recognition respect is that recognition self-respect is what allows you to do your duty from a morally worthy motivation. As was shown in chapter 2, when the moral law weakens a sensible desire merely by delivering a judgment, thereby producing humiliation, it also creates respect, the feeling of awe that moves you to want to defer to the deliverances of your humanity within you.

Is recognition self-respect merely the same feeling as the respect on which we act whenever our actions have moral worth? Phenomenologically, yes; but recognition self-respect can arise not only in response to an immoral sensible desire but also as a result of merely considering one's own humanity. This plays an important role, for by paying attention to one's own dignity, one can resist any tendency one might have to self-abasement.[25]

Self-directed appraisal respect ("appraisal self-respect") plays an important role in projects of moral self-improvement. Kant thinks we have a duty to try to assess our own moral worth so that we can see how far we are from moral perfection, thereby generating in ourselves feelings of humility.[26] Additionally, we should not only determine our degree of moral worth by comparing our conduct to that demanded by the moral law, but we should also compare our moral worth to the moral worth of others, not so that we can feel superior to them, but so that we can receive additional increments of moral motivation by seeing what others are capable of.[27]

Two Kinds of Self-Love

Because there are two kinds of loves, well-wishing and well-liking love, so there are two kinds of self-love, well-wishing and well-liking self-love.

Well-Liking Self-Love

We have well-liking love for others when we delight in their properties (e.g., Newton's intelligence) or in them on the basis of their properties (e.g., having positive feelings for my brother because of his wit). Analogously, we have well-liking for ourselves ("well-liking self-love") if we delight in our own perfections or in ourselves on the basis of our perfections. "The love that takes pleasure in others is the judgment that

we delight in their perfection. But the love that takes pleasure in oneself, a self-love, is an inclination to be well-content with oneself in judging of one's perfection" (*LE-C*, 27:357).

One wrinkle that arises in the concept of well-liking self-love is its relation to relative worth. Do people entertain well-liking self-love because of their own relative worth to themselves? This is perhaps possible; arguably, the suicide who wants to end her life because she expects a future with more pain than pleasure is someone who thinks of life as important only insofar as she can experience sensible pleasure; once she is sure that that pleasure will be swamped by pain, she sees her continued existence as relatively worthless.

More commonly, though, people have well-liking self-love for those perfections of theirs that help them reach their goals (e.g., someone may take pride in her sense of humor) or for themselves on the basis of those perfections that help them achieve their goals (e.g., a person thinks of herself in positive terms because of her managerial skills). It seems as well that a person could have self-directed well-liking for those perfections of theirs that are particularly helpful to *others*. Someone who was wealthy might delight in her wealth because it makes her relatively useful for others. Given Kant's hedonism, it would have to be the case that being relatively useful to others is immediately pleasing[28] for her, which is why she delights in it.

Well-Wishing Self-Love

Well-wishing self-love is easy enough to understand: it leads to actions aimed at promoting one's own happiness; given that this is the kind of behavior that well-wishing

self-love produces, it is no wonder Kant says "there is in all men without restriction a love of well-wishing towards themselves" (*LE-V*, 27:620).

All sorts of motivations can underlie other-directed well-wishing love – I may try to make someone happy because it makes me happy, or because I think it will benefit me in some way, or because it is my job – but this is not obviously the case with well-wishing self-love. After all, while I may try to benefit others as a means to some other end of mine, it would be odd to say that I try to benefit *myself* as a means to some other end.

At the same time, it would also be a little odd to say that I wish myself well because of some positive attitude to myself. As C. S. Lewis asks, when discussing God's command to people to love their neighbors as they love themselves, "how exactly do I love myself? … I have not exactly got a feeling of fondness or affection for myself, and I do not even always enjoy my own society" (Lewis 1996, 105).

I do not think, though, that Kant thinks well-wishing self-love is always motivated by a particular attitude (though he does not rule this out); instead, because of our original predispositions and because of our commitment to the Prudential Maxim, we are designed by nature to desire certain things (e.g., things that conduce to our self-preservation, sexual intercourse, community with others, etc.), and we also judge those things to be good.

Now, we judge the satisfaction of our sensible desires to be good because we expect satisfaction of our sensible desires to produce pleasure, and we like pleasure. But, funny as this question may sound, *why* do we like pleasure?

I think that Kant actually answers this question. Time and again when Kant discusses pleasure, he defines it as "the feeling of the promotion of life" (*LM-L*$_2$, 28:586).[29] Defining pleasure as the feeling of the promotion of life must seem a little obscure, but Kant provides some help in illuminating what this means in the following passage:

> A thing lives if it has a faculty to move itself by choice. Life is thus the faculty for acting according to choice or one's desire. But now this is, practically speaking, the faculty of desire. Since pleasure is thus agreement with the faculty of desire, it is also agreement with life, and displeasure [is] conflict with life. But pleasure and displeasure presuppose sensation. Accordingly I can also say: pleasure and displeasure is a feeling of agreement and conflict or, what is the same, of the promotion or obstruction of life. (*LM-M*, 29:894)

One might think that what Kant has in mind is simple enough: we are living, fragile beings; as such, we need to have some way of knowing when our existence is threatened and when it is promoted. The way this is communicated to us is via sensations, namely pleasure and pain.

I do not think this is right, because for Kant pleasure serves as the ground of every desire. I feel pleasure not only when I eat or sleep, but also when I finish a paper on time or get recognition from my peers. Neither finishing a paper nor getting respect, though, contributes to my continued existence *as a living being*. In what sense do such non-physical pleasures contribute to life?

I think what Kant has in mind is the following. "Life" does not refer primarily to existence in the phenomenal world (though it does include that), but mainly to the causal efficacy of the faculty of desire. As he puts it in the *Critique of Practical Reason*, "**Life** is the faculty of a being to act in accordance with laws of the faculty of desire" (*CPrR*, 5:9n). In other words, if a being acts in accordance with the laws of the faculty of desire, she promotes her life; if she does not, she hinders it (and experiences pain).

The laws of the faculty of desire are not just moral laws (it is not as if we experience pleasure only when we act in accordance with morality); instead, the "laws" of the faculty of desire are determined by a being's predispositions. Each predisposition suggests certain norms of conduct that conduce to life: "To every faculty of the mind one can attribute an *interest*, that is, a principle that contains the condition under which alone its exercise is promoted" (*CPrR*, 5:119). The norms the predisposition to animality brings with it make us judge actions or states of affairs that conduce to animal existence to be good; similarly, the predisposition to humanity makes us judge things that promote human existence to be good; finally, the predisposition to personality makes us judge as good actions or states of affairs in which everyone can act on her animal or human desires without conflicting with anyone else.[30]

So, we like pleasure because we judge things that promote our animal, human, or personal existence—collectively, one's "life"—to be good. Consequently, we wish ourselves well because we see promoting our own lives to be good.

Merit

The last concept we need to clear up before we can start investigating how a person can make herself depraved is that of merit. Kant's account of merit is quite complicated. First, Kant uses "merit" to mean something about a person that can be seen as valuable (as in, "he has a variety of merits"). Second, "merit" can refer to a property of an action (as in, "her action was more meritorious than his"). Third, the merits of a person or an action's merit are supposed to evoke certain kinds of responses, such as respect, esteem, or honor.

Kinds of Merit

In a variety of places Kant uses "merit" to refer to a positive trait of a person. For example, Kant says of "moral self-conceit" (more on this in section 5.2.2) that it "makes an unwarranted pretension to merit. It lays claim to more moral perfections than are due to it" (*LE-C*, 27:357). The morally self-conceited person unjustifiably claims to have merit by claiming herself to have more moral perfections than she actually has. So a "moral perfection"—which could be seen as either a trait that assists the carrying out of moral duties or as one's dedication to leading a moral life[31]—is one example of a merit.

Elsewhere, Kant claims that "envy, in particular, the *invidia qualificata* {aggravated or malignant envy} aims at weakening the other in the possession of his merits", specifically "the destruction of the other's well-being" (*LE-V*, 27:693). So here, a person's well-being seems to be a merit. And earlier, when discussing all three devilish vices, Kant writes that with each of them "we constantly compare ourselves with other men, and feel a chagrin on discovering their good points, whether it be their dutiful conduct, their honor or their well-being" (*LE-V*, 27:692). Here, "good points" seems to be equivalent to "merits"; if so, then merits can be a person's dutiful conduct, her honor, or her well-being.[32]

It seems, then, that merits can be positive traits. Call this sense of merit "trait-merit". There is another sense of merit, though, that applies to action. Call this "action-merit".

To qualify for action-merit, an action must meet one of two conditions: either it must be a wide duty, or it must be performed out of respect for the moral law. In *The Metaphysics of Morals*, Kant writes, "If someone does *more* in the way of duty than he can be constrained by law to do, what he does is *meritorious*" (*MM*, 6:227). The law that

Kant must be thinking of here is juridical law, for it is in fact impossible to do more than the moral law can constrain one to do.[33]

What can juridical law constrain (i.e., coerce) a person to do? Kant gives two senses of what juridical law may constrain:[34] first, there is what is logically possible for juridical law to coerce. In this sense of constraint, a juridical law can coerce a person to do something, but it cannot make a person do something *for a particular reason*. "Duties of virtue cannot be subject to external lawgiving simply because they have to do with an end which … is also a duty. No external lawgiving can bring about someone's setting an end for himself (because this is an internal act of the mind)" (*MM*, 6:239).

The second sense of constraint is what juridical law is *morally permitted* to constrain. So, while it is possible for the state to make people fulfill a wide duty (such as giving to the poor), it is not supposed to, for that goes beyond its proper province. "An authorization to use coercion is connected with any right in the *narrow* sense (*ius strictum*)" (*MM*, 6:233-34).

This discussion shows that when Kant says merit is doing more than can be constrained by law, he means that a person acts meritoriously when she does more than the juridical law is morally permitted to coerce. That is, a person acts meritoriously when she performs a wide duty, and this seems to be true *regardless of whether the motivation that moved her to act is respect for the law or not*.[35] That this is the case can be seen from the fact that Kant in a number of places equates meritorious actions to the performance of wide duties without adding that these actions have to be done from respect to count as meritorious.[36]

Carrying out a wide duty is one way to act meritoriously;[37] another way is discharging a duty, wide or narrow, from an intellectual motive, i.e., from respect for the moral law. "Although there is nothing meritorious in the conformity of one's actions with right (in being an honest human being), the conformity with right of one's maxims of such actions, as duties, that is, respect for right, is *meritorious*" (*MM*, 6:390). This is not to say that the merit of an action is the same thing as its moral worth. An action either does or does not have moral worth – there is no matter of degrees – whereas an action can have more or less merit depending on how hard it is to perform:

> *Subjectively*, the degree to which an action *can be imputed* … has to be assessed by the magnitude of the obstacles that had to be overcome. – The greater the natural obstacles (of sensibility) and the less the moral obstacle (of duty), so much the more merit is to be accounted for a *good deed* (*MM*, 6:228).

It seems that Kant dubs actions meritorious if they show a person to do more than she needs to do. "Meritorious actions include … magnanimity, kindness, etc., since this I cannot require of everyone" (*LE-C*, 27:410).[38] When a person does a wide duty, whether it is motivated by respect or merely by sympathy for others, she does something more than she needs to do by (juridical) law; she did not have to do it, yet she did it anyway.[39] Similarly, when a person does something out of respect, whether that duty is wide or narrow, she again does more than she needed to do. She did not have to perform the action for that reason, yet she did.

Reactions to Merit

Kant writes, "We can value a thing for what it is worth, but high esteem and honor we can give only to that which has merit" (*LE-C*, 27:409). It is not immediately obvious what Kant means when he says that something meritorious (an action or a

person?) deserves high esteem or honor (appraisal respect or mere approval?). We can figure out what he means if we explore his treatments of the "love of honor".

It is clear there is some connection between honor and (action-)merit. "We acquire respect in virtue of good conduct, but honor in virtue of meritorious actions" (*LE-C*, 27:410). Already from this passage we know that when Kant talks about reactions to merit, he means reactions to meritorious people in virtue of their actions.

Honoring someone is the same as approving of or praising her. Kant distinguishes between two ways in which we can love honor (i.e., try to gain approval or praise).[40] We can love honor in the negative sense of trying to avoid others' contempt: "the love of honor is a negative thing; our only concern is not to be an object of contempt" (*LE-C*, 27:408-9); alternatively, we can love honor in the positive sense of trying to gain others' praise or approval: "he who seeks honor, without any ulterior motive, merely in the approval of others, is truly a lover of honor" (*LE-C*, 27:410).[41]

The sense of honor that has most to do with merit is honor in the positive sense, that is, praising or approving of something. Merit, regardless of whether it is motivated by respect or is merely the discharging of a wide duty, is connected to praise because it can be greater or lesser depending on how hard it is for a person to perform an action; the more difficult it is for a person to perform a wide duty or morally worthy action, the more meritoriously she acts,[42] and the more praise she deserves. Praise can vary from what Kant calls "respect"[43] (and what I, to reduce confusion, shall call "juridical respect", i.e., not interfering with or putting down someone), which is granted to actions that are not meritorious, so much as not culpable (i.e., actions that discharge a narrow duty) to "high esteem" or "honor", which refer to actions that have a significant degree of merit.

So far I have discussed the attitudes that action-merit inspires. What about trait-merit? Generally, the attitudes Kant discusses in connection with trait-merit are invidious attitudes, like those motivating devilish vices. However, there is no reason to deny that individual merits can be the object of honor, or love, or even appraisal-respect (at least, if the merit in question is dutifulness).

Just as with worth, merit can be attached to the wrong things. A person can think an evil action is meritorious if it was really hard for her to do (think here of Adolf Eichmann overcoming his natural revulsion at seeing innocent Jews murdered by telling himself that it is his duty to obey the Führer), or she could think that her inner or moral worth depends simply on how meritorious she is, rather than on her dignity or on how closely she comports with the moral law's commands out of respect for them.[44] Indeed, as we shall see, not only are such misunderstandings possible, they happen all the time.

Summing Up the Preliminary Distinctions

So far, there have been lots of terms introduced, as well as distinctions within and relations among them. Because it is necessary to comprehend the precise complexions of these terms to grasp what follows, I present a list of them along with some brief definitions (if in one definition I make mention of another of the defined terms, I boldfaced the defined term):

Inner worth/dignity: An agent's dignity or inner worth is based on her humanity: because she has the capacity to act from respect for the moral law, she has inner worth. Because every agent has humanity, every agent has equal inner worth.

Moral worth: Moral worth is the worth a person has as a result of how she uses her humanity. The closer she is to a perfectly virtuous person, the morally

worthier she is. Actions admit of moral worth, though they either have moral worth or they do not.

Relative worth: A person's relative worth is her usefulness. The more useful she is, the greater her relative worth. Because relative worth has to do with usefulness, a person *A* can have high relative worth for another person *B* and low relative worth for a third person *C*. Actions also have relative worth, and one that comes in degrees depending on their usefulness.

Recognition respect: You show recognition respect for a person if you treat her as an end in herself rather than as a mere means. The feeling that is supposed to give rise to recognition respect is exaltation at a person's dignity, but it could also be the mere desire to obey the (juridical) law. Thus, recognition respect can describe the way we are supposed to treat someone in virtue of her **inner worth**, or it can denote the way we are supposed to feel about her **inner worth**.

Appraisal respect: Appraisal respect is the feeling of moral approval and fellow-feeling we have for someone in virtue of her having **moral worth** or her performing an action that has **moral worth**. Because the degree of a person's moral worth can vary, so can our degree of appraisal respect for her.

Love of well-liking/benevolence: Well-liking love is a feeling of delight we have for a person or thing's perfections, or for a person or thing because of her or its perfections. The perfections that induce us to have this feeling of delight must be ones that we think have **relative worth** for us.

Love of well-wishing/beneficence: Well-wishing love describes actions aiming to benefit someone. A wide variety of attitudes can underlie well-wishing love.

Recognition self-respect: Recognition self-respect mandates that you treat yourself as an end in yourself and never as a mere means, say, to your pleasure. The attitude that underlies recognition self-respect is supposed to be awe and deference to your own **inner worth**.

Appraisal self-respect: You feel appraisal self-respect for yourself on account of your awareness of your own **moral worth**.

Well-liking self-love: This is the delight we feel in our own perfections because of their **relative worth** for us (or even for others), or the delight we feel in ourselves because of our relatively useful (for ourselves or others) perfections.

Trait-merit: Trait-merit refers to the merit someone has because of her individual perfections. Each perfection can be described as *a* merit; when someone praises a person on the basis of her merits, he praises her because of what (he assumes) she had to do to acquire those traits, or because of their usefulness for her.

Action-merit: An action has action-merit if it is good to do but is also not coercible (either logically or morally). Wide duties undertaken out of respect for the moral law or non-deplorable motivations all have action-merit. Narrow duties do not have action-merit, unless they are motivated by respect for the moral law.

Honor: Honor is praise or approval. We honor **action-merit** predominantly, but we can also honor **trait-merit**. The greater the action-merit, the greater the honor it deserves.

I can now present Kant's explanation of the possibility of willful evil.

4. Willful Evil

Kant's explanation is as follows. People are capable of acting on depraved maxims – i.e., maxims according to which some sensible desire presents a better reason for action than a moral obligation – because they can tell themselves stories, or entertain fantasies, according to which the normal force of the moral law is suspended, either for them, or members of their caste, or against persons of such-and-such a kind, or for actions of this type, or just this one time, etc.

People can come to believe these stories by engaging in self-deceptive techniques of rationalization or conscience-avoidance. After enough use of the techniques, a fantasy can seem to a person to be true, although there is some part of her somewhere that does not believe it.

One way in which a person can downgrade a moral obligation even though she knows it is supposed to be overriding is by *mostly* believing a story. Because of the story, one or more of her sensible desires looks to be more important than her obligation, but because a part of her does not fully believe the fantasy, she also knows the obligation to have practical paramountcy. Such a person can know that she should not do what she is about to do, but then distract her attention back to what she, according to the fiction, is permitted (or even obligated) to do.

Another way in which a person can act on a depraved maxim is by so habituating herself to its permissibility or its false obligingness that she ends up not knowing, on any level, that it is wrong. A person does not act willfully evilly by acting on that maxim, though she does act evilly (and not out of frailty or impurely); rather, she acts willfully evilly at the point where she realizes she is bringing this habit upon herself but presses forward.

These two possible ways of acting on depraved maxims are open to the following objection: "all you've said so far is that if a person convinces herself that acting on some sensible desire is more important than carrying out some competing moral obligation, she can act willfully evilly by acting on the maxim according to which that sensible desire is of greater importance than the moral obligation. But acting on *that* maxim isn't willfully evil; it's just confused or ignorant. Such a person simply doesn't *know* that the moral law is overriding – if she did, then she would either violate it out of frailty, or she would simply satisfy her moral obligation. Thus, there is *no way* for a person to say, ***all at the same time***: I know *x* is my obligation; I know I have more reason to do *x* than ~*x*; yet I have overriding reason to do ~*x*; therefore I shall do ~*x*."

To *almost* all of this I say: yes, that is right. There are, however, two issues where I differ. First, if a person makes herself into the kind of person who comes to think of certain obligations (ones she used to think of as overriding) as ones she no longer has very strong reason to discharge, and so ignores them, then this is evil, even if it is not willful evil. Second, there is a sense in which a person knows, all at the same time, that *x* provides her with an obligation that overrides ~*x* and that ~*x* provides her with an obligation that overrides *x*. The sense is this: one "part" of her knows that she ought to discharge her duty, while another "part" "knows" that discharging her duty is less important than carrying out her obligation. Since she is composed of both these parts, she knows this at the same time, though she cannot accept both in the same thought, as it were. This is how self-deception works for Kant.[45]

This answer is not likely to satisfy the objector, for she could just level this criticism: "your answer here does no good; rather, it just pushes the mystery back one

step. Either the agent is morally responsible for deceiving herself – for choosing not to entertain the thought of her moral obligation's bindingness at the same time she thinks her sensible desire has more going for it – or she is not. If she is not morally responsible for her self-deception, then she is not evil. If she is morally responsible, we have to ask how she can be. Any answer you give will either be similar to the one you just gave (she doesn't allow herself to confront her self-deceptiveness), in which case that is just pushing the question back; or you have to claim that she consciously deceives herself, even though she knows it is wrong, because she also thinks it is right; which is impossible. So wholehearted or willful evil, or wickedness is impossible."

(Note that the critic could also say the same thing about frailty, for frailty involves self-deceptively convincing yourself that you cannot overcome your own sensible desires. But since you can, and since you know this (if you did not know this, you would not be blameworthy for your *akrasia*), there is no way to make sense of how someone can act frailly and culpably at the same time.)

There is, I think, only one route for Kant to go at this point, and it has to do with the evil *Gesinnung*. There are, I think, two senses of depravity that Kant uses, though they are structurally similar in that both subordinate the moral law to a sensible desire. The first sense of depravity is the one Kant mentions in the *Religion*, as the third grade of the propensity to evil *qua* tendency. This kind of depravity afflicts particular disposition-maxims, and is the endpoint of a process beginning with frailty and going through impurity.[46] A person could be frail regarding disposition-maxims of one sort (say, having to do with food or sex), impure with regard to disposition-maxims of another sort (say,

having to do with her career), and depraved in this sense as regards disposition-maxims of a third sort (say, having to do with supporting an immoral public figure).

Call that sense of depravity "surface depravity". There is another sense of depravity, "deep depravity", which is the sense that characterizes a person's initial acquisition of her evil *Gesinnung*. As stated in chapter 3, a person activates her susceptibility to evil through a free choice, and because this choice generates in her a tendency to evil that is itself evil, and because, in addition, this choice is morally forbidden, a person who makes it shows herself to be (deeply) depraved.

This deep depravity is not initially apparent (though the tendency to evil that is its fruit serves as evidence of it, at least once one knows Kant's theory). However, as a person grows up in an evil society, she comes closer and closer to manifesting surface depravity; when she finally performs a depraved action, deep depravity reaches its fullest expression. This is because *deep depravity is what directly enables surface depravity*.

Deep depravity enables surface depravity by allowing a person to see her self-deceptive techniques – rationalization and avoidance – as having more going for them, all things considered, than obeying the moral law. This is how willful evil is possible – *the self-deception is the willful evil*, which "transfers" its evil to all the evil actions it enables; moreover, the inaugural evil decision to activate one's susceptibility to evil – which enables self-deception for which one can be fully morally responsible – is also willful evil.

This is quite speculative, and I doubt that it was Kant's actual view, but this is because I do not think he spent much time trying to figure out self-deception. That is, I am not sure Kant *had* a view, or at least a worked-out view, on the issue of self-

deception. However, I think it comports very well with Kant's view, and allows us to explain remarks such as this: "The greatest violation of a human being's duty to himself regarded merely as a moral being (the humanity in his own person) is the contrary of truthfulness, *lying*" (*MM*, 6:429) and:

> It is noteworthy that the Bible dates the first crime, through which evil entered the world … from the first *lie* … and calls the author of all evil a liar from the beginning and the father of lies. However, reason can assign no further ground for the human propensity to *hypocrisy* … although this propensity must have been present before the lie (*MM*, 6:431).

In what follows, I shall explore the nature of surface depravity in somewhat greater detail, mainly by focusing on the kinds of stories or fantasies people can come to believe about themselves, and the means by which they come to believe them.

5. Evil People

In both the Collins lectures on ethics (1784-85) and the Vigilantius lecture notes (1793-94) as well as the *Critique of Practical Reason* Kant presents a fascinating variety of evil people. Each variety of evil person has a different conception of herself and her relation to the moral law, and in each case her self-conception is associated with a false standard of valuation. This standard of valuation might make relative worth the most important determinant of a person's total value, or it might allow for differing degrees of inner worth, or a sense of moral worth not dependent on the actions a person has performed; all these are possible. Kant's examples do not exhaust his complicated network of ideas regarding ways of valuing, the sources of value, and the kinds of responses to value that are possible, but they do show there to be a lot of different ways of going wrong.

Before going on, one might wonder why it is permissible for me to use the Collins lecture notes. After all, in those notes, as was shown in last chapter, Kant denies a

propensity to evil, limiting the kinds of evil action to ones stemming from frailty and ignorance. It should be remembered, though, that he explored devilish vices, despite thinking them to be impossible. So, why not go even further and explore evil self-conceptions that are, at the end of the day, not really possible? Thus, I take it that in the Collins lecture notes Kant made the curious decision to explore a variety of evils and evil ways of thinking, even though he knew he would have to claim that matters could not be (for theological reasons) the way they appeared to us.

5.1. Solipsism

The first kind of evil is solipsism. Someone acts solipsistically when she engages in an excess of well-wishing love towards herself (i.e., when she undertakes to perform too many actions that aim at the promotion of her own happiness). "The *love of well-wishing towards oneself;* considered exclusively in relation to oneself, and thus without any regard for the duty of love towards others, this is solipsism or egoism – the {loving self-approval}" (*LE-V*, 27:620).

The solipsist tries to bring about too much happiness for herself; she has an excess of well-wishing self-love. Remember that well-wishing self-love, as opposed to well-wishing other-love, cannot have just any motivation moving it. Whereas I could wish you well so that you will help me in the future, or because it is my duty, or merely because I get a kick out of it, I cannot wish myself well for just any reason (say, because I hope I will help me in the future). Rather, I wish myself well in the first place – I seek personal happiness – because I have predispositions to animality and humanity that make me see actions that harmonize with my animal or human life as a good thing to do.

And yet Kant seems to indicate that the basis of solipsistic behavior is not merely acting according to the norms of one's animal or human predispositions a little too often (i.e., sometimes at the expense of one's duties); rather, he sees "loving self-approval" (*amor acquiescentae in semet ipso*) at its root. Moreover, solipsistic behavior infringes especially on one kind of behavior: one's imperfect duty of beneficence to others. "The *love of well-wishing towards oneself;* considered exclusively in relation to oneself, and thus without any regard for the duty of love towards others, this is solipsism or egoism"; or, as he later notes:

> The contradictory opposite of philanthropy is the *animus frigidus*, or callousness, which consists in indifference towards the state of other people – a person who is devoid of love for others. This indifference rests on s*olipsismus*, or a care for one's own welfare that is guided solely by partiality for oneself. (*LE-V*, 27:672)

So, a person acts solipsistically when her loving self-approval moves her to direct a disproportionate amount of her energies away from aiding others (or at least, that could have been used to aid others) to the promotion of her own happiness. What is "loving self-approval", though? Well, "approval", as already mentioned, has to do with honor or praise, and honor/praise is to the proper response to action-merit. According to this line of thinking, then, the solipsist tries to promote her own happiness, at the expense of others' happiness, because she finds herself action-meritorious.

There is one problem with this interpretation, however. In Kant's examples, action-merit results either from performing a wide duty of beneficence to others, or from performing a narrow duty out of respect. What makes a person solipsistic, though, is that she does not perform her wide duties of beneficence for others, instead focusing on herself. But if she does not help others, how can she approve of herself? Perhaps she

honors herself because she performs narrow duties out of respect; or perhaps she deceives herself into thinking she helps others but actually does not.

I have a different theory. I think we should pay attention to the "loving" part of "loving self-approval". Love comes in two flavors, of course, well-wishing and well-liking, but we are trying to explain why someone is disproportionately well-wishing self-loving; so it will not do to explain that by adverting to well-wishing self-loving. Consequently, I take it that someone is solipsistic if, because of her well-liking love for herself (her indirect delight in herself on the basis of her perfections, or her direct delight in her perfections) she seeks to make herself happy beyond what she is morally permitted to do.

We can tie this approval and merit if we imagine that the perfections in which she delights are merits – any old trait that has a value for people. Overall, then, my feeling is the solipsist is someone who delights in herself on the basis of her merits (her well-being, her dutifulness, her skills, her social standing – it could be just about anything that has relative worth), and so tries to make herself disproportionately happy. (Note: it seems to me both that someone could be a solipsist, and that someone who is not a solipsist – say, an "adequacy fantasist" (a species of evil person I articulate in the next subsection of section 5) – could act solipsistically on occasion without solipsism being her main self-conception.)

Let me note a pair of things about solipsism. First, although solipsists subordinate their wide duties of beneficence to their desire for their own happiness, it does not follow that they will subordinate any perfect duties to their own happiness. It is one thing to ignore your charitable duties, and quite another to murder or deceive someone.

Second, I am assuming that anyone who is a solipsist is willful about her solipsism, or comfortable about ignoring her imperfect duties to others. It is not that she thinks she should help others more than she helps herself but just finds it too difficult; rather, making others happier is either something that just never occurs to her because of her self-preoccupation, or it is something that simply seems to be of less importance than indulging her own desires.

Let us explore this second truth about solipsism, for this is where the evil of solipsism comes in. The solipsist who finds her own happiness to be obviously more important than helping others, and who, consequently, does not intend to make any sacrifices of personal happiness on others' behalf,[47] makes it her principle to ignore her imperfect duties to others. This is not just a lack of merit, but is out-and-out culpable:

> Fulfillment of {imperfect duties} is *merit* (*meritum*) = +a; but failure to fulfill them is not in itself *culpability* (*demeritum*) = -a; but rather mere *deficiency in moral worth* = 0, unless the subject should make it his principle not to comply with such duties. … Every action contrary to duty is called a *transgression* (*peccatum*). It is when an intentional transgression has become a principle that it is properly called a *vice* (*vitium*). (*MM*, 6:390)

A solipsist, then, thinks it okay to ignore her imperfect duties to others to help herself. (And insofar as one reason to perfect yourself is to make yourself capable of promoting others' happiness, she sees no reason to do this either.[48]) Why does she think this? As stated above, she thinks this because she takes such delight in her own perfections, or because she has such praise for her own "merits"—whether this is her well-being, her status, her (self-supposed) dutifulness, or whatever—that she thinks it is *more important* or *better* to make herself happier than to help others. But insofar as she judges this to be the case, she expresses her commitment to the principle: the happiness

of people who excel in trait-merit is significantly more important than the happiness of people who do not so excel.

We can ask why the solipsist accepts such a principle. Any reason for why she accepts such a principle will itself be grounded in another principle, but it is worth spelling out the various reasons anyway. First, a solipsist could think that the most important kind of worth is relative worth; the more relative worth a person has – and this will be determined on the basis of her merits, including such things as her usefulness to others – the more important she is, and the stronger our obligations to her to make her happier.[49]

As stated so far, this principle suggests that if we have to choose between violating a perfect duty to someone of low relative worth and making someone of high relative worth happy, we should do the latter. Someone who is evil *only* insofar as she is a solipsist, though, will not think *this*. She will still be constrained, by her recognition respect for people's inner worth, from going against their dignity. To the extent, though, that we are supposed to contribute to others' happiness because of respect for their inner worth, she prioritizes relative worth over inner worth. And this is evil.

Alternatively, someone could be a solipsist simply because she thinks she has more moral worth than others. Such a person would think that the most important kind of worth is moral worth; she might conclude that, the more moral worth a person has, the stronger are our obligations to make her happier. Someone being a solipsist for this reason, however, is not very likely, because she would have to think of moral worth in a way where it has nothing to do with helping others, but only to do with keeping to one's perfect duties and making people of high moral worth happy. While it is perfectly

coherent to adhere to some standard according to which moral worth has nothing to do with helping others, such a standard would have to be quite warped, and it is hard to think of one that results *just* in solipsistic neglect of one's duties to others without also leading to situations in which one violates perfect duties to others as well.[50] This is not to say, though, that there is no such principle.

I would like to enter two observations about solipsism before going on to the next type of evil person. First, while a solipsist who justifies her disproportionate attention to her own happiness may be logically committed to devoting a similar amount of attention to others she regards to have equal or higher relative worth, it does not follow from this logical commitment that she will actually do this. She could violate her "duties" of aid to other similarly gifted people. She could do this for one of two reasons. Either she could be quite delusional, and literally think that she has the highest amount of relative worth of anyone out there; or she could adhere to the principle as I have stated it, but not live up to it out of a kind of weakness of will. *Akrasia* of this second sort would not be surprising to find in a solipsist, because on Kant's view the main reason she is solipsistic in the first place is that she wants to justify to herself courses of action that allow her to be as happy as possible without having to live up to the burdens of morality (more on this later). Consequently, she would often feel pressure to violate those "duties" that even her relatively laissez-faire standard of conduct moves her to endorse.

My second observation is a connection of solipsism to my discussion of pleasure and life from section 3.3.2.2. There, I pointed out that what counts as pleasurable to a person is that which "promotes her life." Acting in accordance with the norms of one's predispositions to animality and humanity gives rise to the feeling of life-promotion, i.e.,

pleasure. I think, though, that something similar is going on with solipsism (and all other evil self-conceptions): because she adheres to a standard in which (let us say) people of high relative worth deserve more happiness than people of lower relative worth, things that conduce to her own happiness will seem to promote her life more than actions she takes to promote others' happiness. Such other-helping actions simply will not give her as much pleasure as helping herself, because her standard of evaluation judges actions of that kind not to promote her life.

Solipsism is, essentially, a self-conception that justifies what (from the point of view of morality) is an excess of well-wishing self-love. I find no indication that Kant thinks that people who are evil exclusively because they are solipsistic are very common. On the other hand, he seems to think that evil people of the sort that I call "adequacy fantasists" are all over the place.

5.2. Moral Fantasies

In the Collins lectures on ethics, Kant describes two kinds of "moral fantasies":

> Moral fantasies may relate either to the moral law itself, or to our moral actions. The first such delusion is to fancy of the moral law that it is indulgent in regard to ourselves. But the other is to fancy of our moral perfections that they are in conformity with the moral law. The first is more harmful than the second, for if a man fancies that his perfections are compatible with the moral law, it is still easy to dissuade him of this, by pointing to the purity of the moral law. But if a man frames for himself the idea of an indulgent moral law, he has a false law, whereby he also creates maxims and principles such that even his actions can then have no moral goodness. (*LE-C*, 27:348)

The first fantasy is "to fancy of the moral law that it is indulgent in regard to ourselves"; in other words, a fantasizer of this first kind thinks that the moral law, though issuing commands of overriding normative authority, happens to make a special dispensation for her such that she does not have to meet the same commands she thinks everyone else has to meet. The second kind of fantasy is "to fancy of our moral perfections that they are in conformity with the moral law." As Kant presents things, a person who entertains this

second moral fantasy really accepts the moral law, but defectively thinks that she is doing just fine by it.

I call a moral fantasist of the first sort an "exceptionalist fantasist" (her fantasy is, correspondingly, the "exceptionalist fantasy") and one of the second kind an "adequacy fantasist". I shall explore the adequacy fantasist before the exceptionalist fantasist, because I think that exceptionalist fantasies arise, in large part, from dynamics and tensions internal to the adequacy fantasy.

5.2.1. The Adequacy Fantasy

Despite Kant's language in *LE-C*, 27:348, I do not think that even he thinks the adequacy fantasist really accepts the moral law; language he uses to describe the adequacy fantasist in later parts of the Collins lecture notes show this to be the case. However, the adequacy fantasist accepts a fundamental principle much closer to the moral law than the exceptionalist fantasist (though she, like the exceptionalist, is still evil).

The first question we must ask of the adequacy fantasist is why she thinks her perfections, such as they are, are in line with the moral law's demands. In an important passage illuminating both the adequacy and exceptionalist fantasies (but which, owing to its length, I divide into two parts), Kant explains:

> The love that takes pleasure in others is the judgment that we delight in their perfection. But the love that takes pleasure in oneself, a self-love, is an inclination to be well-content with oneself in judging of one's perfection. *Philautia*, or moral self-love, is to be contrasted with arrogance, or moral self-conceit. The difference between them is that the former is only an inclination to be content with one's perfections, whereas the latter makes an unwarranted pretension to merit. It lays claim to more moral perfections than are due to it; but self-love makes no demands, it is always merely content with itself and devoid of self-reproach. The one is proud of its moral perfections, the other is not, believing itself merely to be blameless and without fault. *Arrogantia* is thus a far more damaging defect. (*LE-C*, 27:357)

Interestingly, Kant seems to equate what I call the adequacy fantasy with something he calls "*philautia*" or "moral self-love". Moral self-love is "an inclination to be content with one's perfections" and "is always merely content with itself and devoid of self-reproach"; the moral self-lover is not proud of her perfections, but does believe herself "to be blameless and without fault." (This is enough, I think, to justify the claim that the adequacy fantasist, who fancies that her moral perfections are in conformity with the law, is the same as the moral self-lover, who is content with her perfections.)

The adequacy fantasist adheres to a standard according to which she is no moral saint, but is also not morally deficient. By her own lights, she is doing well enough, although she could be doing better.

The problem with the adequacy fantasist is that she derives her ideas of duty from her idea of merit. As I pointed out in section 3.4 of this chapter, action-merit attaches to the performance of wide duties (if done from an acceptable motivation) and also to the performance of narrow duties if they are done from respect. The adequacy fantasist fulfills her narrow duties, as juridically understood, but does not really go for wide duties (though she may perform one here or there on occasion). So, she thinks she carries out what she owes to society; even if she does nothing particularly meritorious, she still does nothing culpable either.

There are two immediate problems with the adequacy fantasy. One springs to mind upon a cursory overview of an adequacy fantasist. First, while it is true that from the point of view of the state, a person needs only to fulfill her narrow obligations, from the point of view of morality, a person is supposed to fulfill her wide obligations:

> it is indeed <u>impossible</u> that in the sight of God, <u>as the law of the highest morality</u>, we can do more than is incumbent, since in regard to Him, <u>everything is required</u>; but in relation

> to other men, we can certainly have merits, if we measure our actions against our coercive duties, e.g., beneficence towards the poor. (*LE-V*, 27:665)

Note that in God's eyes, everything is *required*. That is, to truly discharge her duties a person is supposed to devote *all* her efforts to self-perfection and aid to others, at least so long as she violates no perfect duties in doing so.[51] So, by thinking of herself as morally blameless merely because she does not violate any perfect duties, she shows herself to misconceive of morality.

The second problem with the adequacy fantasy is revealed when we look at the rest of the quotation from *LE-C*, 27:357:

> *Philautia* tests itself against the moral law, not as a guiding-principle but by way of examples, and then one may well have cause to be self-satisfied. The examples of moral men are standards drawn from experience; the moral law, however, is a standard set by reason; if the first of these is used, the result is either *philautia* or *arrogantia*. The latter arises if the moral law is thought of in a narrow and indulgent fashion, or if the moral judge within us is partisan. The less strictly the moral law is taken, and the less strictly the inner judge passes judgment upon us, the more arrogant we tend to be. Self-love differs from esteem. The latter refers to inner worth, love to the relationship of my worth in regard to well-being. (*LE-C*, 27:357)

The reason the adequacy fantasist thinks of herself as fine, even though she does not really address her imperfect duties, is that she gets her sense of moral worth, not by comparing herself directly to the moral law, but rather by comparing herself to the moral worth of others, as exhibited in their behavior.

The adequacy fantasist thus adheres to the principle: a person's moral worth should be assessed not through comparison to the moral law, but by comparison to her fellows. If she has less moral worth than her fellows, she is culpable; if she has more or less the same amount, she is blameless, and if she has more, she has cause to be self-satisfied.

Unfortunately, adhering to such a principle is not only evil, but extremely dangerous. For one thing, what your fellows take to be a measure of moral worth may not, in fact, be a measure of moral worth but might be positively immoral:

> Men like, in general, to have examples, and if one exists they are happy to excuse themselves, on the ground that everybody lives that way … A bad example, however, is a stumbling-block and gives occasion for two evils; for imitation as a pattern, and for excuse. (*LE-C*, 27:334)

Thus, if everyone takes it for granted that it is okay to lie to advance your career, then you can freely lie, and feel no guilt about it.

Another problem is that a person who takes his cues about his moral worth from the example of others does not have much reason to improve himself if he thinks that he is doing fine: "if we have examples of moral imperfection before us, we can flatter ourselves at our own moral imperfection" (*LE-C*, 27:294).

Finally, a person who adheres to the principle that the proper way to assess a person's moral worth is by seeing where he stands in relation to others opens himself up to the possibility of becoming self-conceited and an exceptionalist fantasist (which two are not necessarily the same thing):

> moral self-esteem, which is founded on the worth of humanity, must never be based on a comparison with others, but only on comparison with the moral law itself. People are very much inclined to take others as the measure of their own moral worth, and <u>if they then believe themselves superior to some, it would be self-conceit to think thus</u>; but the latter is far greater if one believes oneself to be perfect in comparison with the moral law. (*LE-C*, 27:349)

The danger here is clear: if I determine my own moral worth in comparison to others, I may conclude that I am morally worthier than someone else; and if I do that, I can become self-conceited or arrogant. Kant is not saying that everyone has equal moral worth – Kant is clear not only that they do not, but that real disparities in moral worth can sometimes even be noticed[52] – but rather that without comparing oneself to the moral

law, which acts as a check on a person's self-satisfaction by always humiliating her to some degree,[53] a person can end up thinking of herself as morally superior to others, and so deserving of some special rights and privileges.[54] And at that point, she becomes liable to the exceptionalist fantasy.

5.2.2. The Exceptionalist Fantasy

The exceptionalist thinks that the moral law indulges her. Presumably, that means that the moral law allows her to do things that it would not do for others. But there are at least two ways in which a moral law (not: The Moral Law) could be indulgent.

First, the law could simply run, "a person's moral worth should be assessed not through comparison to the moral law, but by comparison to her fellows." This is the same standard of evaluation the adequacy fantasist uses; what separates the adequacy fantasist from the exceptionalist fantasist is that the latter believes not only that her moral worth should be determined through comparison to other people, but also that she comes out well ahead of her peers when this standard of comparison is used.

Now, her belief that she has more moral worth than her peers need not be true;[55] as Kant says, the self-conceited person "makes an <u>unwarranted</u> pretension to merit.[56] {She} lays claim to more moral perfections than are due to {her}" (*LE-C*, 27:357). What distinguishes this form of moral self-conceit (moral exceptionalism) is that the conceited person carries herself as though she is better than her peers *because of her allegedly greater moral worth*.

There is another form of the exceptionalist fantasy, though, one that uses a different standard from that of the adequacy fantasist. According to this form of the exceptionalist fantasy, the law indulges the exceptionalist, not in the sense that it

describes her as better than others because of her accomplishments; rather, it describes her as better than others because of some property of hers that has nothing to do with accomplishments at all. On this version of exceptionalism, the exceptionalist lays claim to greater *inner* worth; the moral obligations people have differ depending on their inner worth: "*arrogantia* is pride, when we presume to a value that we do not possess; but if we lay claim to precedence over others, that is haughtiness; in that case, we put down the other, and deem him lesser and lower than we are" (*LE-C*, 27:457). Those of high inner worth (like the exceptionalist herself) have obligations to each other, and may or may not have obligations to those of lower worth; regardless, the structure of obligations between people of putatively lower and higher inner worth is an inegalitarian one.

The standard of evaluation to which this kind of exceptionalist commits herself thus runs: one ought to have "high" recognition respect for people of high inner worth, and "low" recognition respect for people of low inner worth. There are a variety of ways such a principle could be applied; it could command deontological constraints on how everyone treats everyone else, where the nature of those constraints depends on the kind of inner worth one has (presumably, if you have lower inner worth, then the constraints someone of higher inner worth has towards you can be overridden depending on the circumstances). Alternately, it may be that only people of lower worth have obligations (namely, to those of higher worth), without having any rights. Or, inner worth could come in a wide variety of degrees, never depending on anyone's merit or relative worth, but only on her (say) alleged divinity.

5.3. Despondency

So far I have discussed kinds of evil that depend on supplanting the moral law with another law in its stead, a law that is more permissive to the evil person whose law it is. However, there is another kind of evil, one where the evil person accepts that the moral law is the supreme authority, but (falsely) believes that she cannot live up to that authority. This kind of evil Kant calls, alternatively, "timorousness", "despondency", and "pusillanimity":

> humility can … have injurious consequences, if it is wrongly understood. For it brings timorousness and not courage with it, if a man believes that owing to the defectiveness of his actions they never comply with the moral law, from which inertia arises thereafter, in that he ventures to do nothing at all. (*LE-C*, 27:350)[57]

A person becomes timorous if she concludes that, because of repeated humiliations – instances where she compares her moral worth to what the moral law commands of her and finds herself wanting – there is no point in trying to be moral; if it is impossible to ever fully succeed, so why try? Interestingly, Kant seems to agree with the despondent person, at least to the extent that he thinks that fully satisfying the moral law's demands are impossible on one's own: "To remedy … timorousness, bear in mind that we may have hope that our weakness and frailty will receive a supplement through divine aid, if we have but done as much as we were able to, knowing our capacity" (*LE-C*, 27:350-51). In other words, a person cannot fully live up to the moral law (this seems to be true even if she tries as hard as she is able), but can meet its requirement if God supplements her effort in the right way.

Timorousness is evil first of all because of the attitude underlying it; somewhat childishly, the despondent person concludes that if she cannot be perfectly moral, she might as well not try at all. But because she does not try, she will be evil in a second way, namely by failing to live up to her moral obligations. She may still seek after some non-

moral merit (say, trying hard to become excellent at some non-moral endeavor), and she will still cling to the notion that she has relative worth. However, it is not entirely clear that she will believe that she has inner worth. What gives a person inner worth, after all, is the ability to comport herself in a certain way simply because the moral law requires her to. If a person thinks, though, that she (sometimes, anyway) cannot do this, then inner worth loses (at least some of) its luster.

6. Becoming Evil

I have outlined four different kinds of evil self-conception: solipsism, the adequacy fantasy, the exceptionalist fantasy, and timorousness. What each of them has in common is a fundamental standard of evaluation different from the moral law, according to which things of a certain kind count as pleasant or unpleasant. What I have not yet discussed, though, is how a person assumes one of these evil self-conceptions and how her adherence to an evil *Gesinnung*, in combination with her tendency to evil, encourages a person to adopt worse and worse fundamental modes of assessment.

6.1. Conscience

Conscience is "practical reason holding the human being's duty before him for his acquittal or condemnation in every case that comes under a law" (*MM*, 6:400).[58] That is, when a person judges that an action she is about to take, is taking, or has already taken, is morally right or wrong, and so feels good or bad about herself as a result, she uses her conscience. Conscience thus involves the understanding and moral feeling:

> We have a faculty of judging whether a thing is right or wrong, and this applies no less to our own actions than to those of others. This faculty resides in the understanding. We also have a faculty of liking and disliking, to judge concerning ourselves, no less than others, what is pleasing or displeasing there, and this is the moral feeling. Now if we have presupposed the moral judgment, we find, in the third place, an instinct, an involuntary and irresistible drive in our nature, which compels us to judge with the force of law concerning our actions, in such a way that it conveys to us an inner pain at evil actions, and an inner joy at good ones, according to the relationship that the action bears to the

> law. (*LE-C*, 27:296-97)

Conscience is not a mere label for those occasions when a person engages her faculty of understanding and of moral feeling at the same time. Rather, it is itself an instinctually applied faculty, that is, a faculty that forces a person to apply her understanding of moral laws to her own situation, and to respond with moral feeling. Conscience:

> is not a mere faculty, but an instinct, not to pass judgment on, but to direct oneself. We have a faculty of judging ourselves according to moral laws. But of this we can make use as we please. Conscience, however, has a driving force, to summon us against our will before the judgment-seat, in regard to the lawfulness of our actions. *It is thus an instinct, and not merely a faculty of judgment*. (*LE-C*, 27:351)

A person's conscience does not judge her only after the fact; it also examines her both before and during the commission of an action. When a person uses conscience to investigate the nature of an action before she undertakes it, her conscience is an "examining conscience"; while vetting her action as she carries it out, it is an "accompanying conscience"; and after a person has completed an action, her conscience is a "judging conscience".[59]

One might get the impression from the fact that a person's conscience can judge her before, during, and after the execution of an action that it is always judging her. This is not what Kant thinks, though.

First, Kant claims that people need to train their consciences. There are two ways in which a person must train her conscience: first, she has to learn how to apply the moral law to her particular case, and second, once she has some aptitude in doing this, she must also become proficient at determining the degree of her moral responsibility for an action (*LE-V*, 27:576).

In addition to training how conscience to undertake these two tasks skillfully, a person has to learn to attend to the voice of conscience within her. Kant calls a person's

readiness to pay attention to her conscience "conscientiousness": "a conscience consists in the ability to impute one's own *factum* to oneself, through the law itself, and the readiness to do this is conscientiousness" (*LE-V*, 27:575); a person, then, needs both to train her conscience and to become conscientious.[60]

So, even if Kant were to hold the view that conscience always speaks before and during the commission of an act, he explicitly denies that its voice is always intelligible or audible. But Kant is not committed to the view that conscience always speaks before someone performs an action. This is because people can, at least in regard to certain kinds of action, silence the voice of conscience in them completely:

> The *conscientia concomitans*, or accompanying conscience, at length becomes weak through habituation, and in the end one becomes as accustomed to vice as to tobacco-smoke. Conscience eventually loses all respect, and then, too, the accusation ceases, having become superfluous, since nothing is any longer decided or carried out in the courtroom. (*LE-C*, 27:356)

Thus, if someone indulges a vice often enough, her conscience eventually stops accusing her altogether regarding it.[61] Correspondingly, a person who completely overcame her temptation to violate a particular moral law (if such a person could ever exist) would also have a quiet conscience when it came to following that law: "if a rational creature could ever reach the stage of thoroughly *liking* to fulfill all moral laws, this would mean that there would not be in him even the possibility of a desire that would provoke him to deviate from them" (*CPrR*, 5:83-84).

But how is it that a person can ignore conscience in those cases where its voice is audible and intelligible? Kant describes two kinds of techniques a person can use to defy her conscience. First, she can avoid its commands:

> we cannot practice this {moral} law so purely; our actions are very imperfect by comparison, so that they are even blameworthy in our own eyes, if only we do not stifle our inner tribunal, which judges according to this law. Anyone noticing this would

> eventually have to give up observing such a law, since he could not face up to so holy and just a tribunal, that judges by this law. (*LE-C*, 27:317)

Here Kant points out that a person, unless she is constantly struggling to be virtuous, cannot abide her conscience constantly needling her to do better. Consequently, she eventually grows tired of her interfering conscience, and like a rebellious teenager with her parent, she simply stops paying attention to her conscience's finger-wagging.

Alternatively, a person can try to rationalize to herself (i.e., to her conscience) why the moral law's commands do (or did) not apply to her in the case before the court:

> A human being may use what art he will to paint some unlawful conduct he remembers as an unintentional fault, – as a mere oversight which one can never avoid altogether, and so as something in which he was carried away by the stream of natural necessity – and to declare himself innocent of it … he *explains* his misconduct by certain bad habits, which by gradual neglect of attention he has allowed to grow in him to such a degree that he can regard his misconduct as their natural consequence. (*CPrR*, 5:98)

6.2. Corruption

As was mentioned in section 4 of this chapter, everyone starts out evil, even when she is not depraved in regards to any particular kind of maxim:

> We cannot start out in the ethical training of our connatural moral predisposition to the good with an innocence which is natural to us but must rather begin from the presupposition of a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition (*Rel*, 6:51).

Although everyone starts out evil, they become depraved in regard to specific maxims either through the education they receive or through opportunities for self-gratification that present themselves.

Kant complains that moral education often miseducates children by directing them to strive after performing meritorious actions:

> I do wish that educators would spare their pupils examples of so-called noble (supermeritorious) actions, with which our sentimental writings so abound, and would expose them all only to duty and to the worth that a human being can and must give himself in his own eyes by consciousness of not having transgressed it; for, whatever runs up into empty wishes and longings for inaccessible perfection produces mere heroes of romance who, while they pride themselves on their feeling for extravagant greatness,

> release themselves in return from the observance of common and everyday obligation, which then seems to them insignificant and petty. (*CPrR*, 5:155).

There are two problems that can result from emphasizing merit in the moral education of children. First, by training children to attend to merit, one makes it less likely that they will think about acting from respect. Consequently, they worry about performing wide duties, not because it is the thing to do, but because such actions will make them appear relatively, or even morally, worthy in the eyes of their peers. This moves children to make their main standard of evaluation others' judgments of their relative worth – i.e., it influences them to become adequacy fantasists. (Indeed, by moving them to aim first of all at meritorious actions, especially when those actions are difficult, it could simply set them up for disappointment when they fail, perhaps pushing them to despondency.)

The second problem that arises from this kind of education is that children think that acting in accord with perfect duty, insofar as it requires little in the way of effort, has nothing of moral significance to it. While it is true that wide duties carried out in the face of great dangers can be both morally worthy and quite meritorious, and are indeed the kinds of actions most elevating to us because they show us the reaches people are capable of thanks to virtue,[62] it is all too easy to focus on the merit of such an action at the expense of duty. One who focuses in this way will take the wrong lesson from her education (and it is the wrong lesson even if the teacher thinks it is the right lesson) and will conclude that a person of great (action- or trait-) merit is exempt from the duties that govern others. In this way, an exceptionalist fantasist can be created.

Education is not the only way to produce moral fantasists. People can also end up in that boat if they react in the wrong way to the moral law. A person who sincerely tries to live up to the moral law, but who nonetheless fails to do as much as she knows she is

supposed to, will find herself humiliated by her conscience. “True humility follows unavoidably from our sincere and exact comparison of ourselves with the moral law (its holiness and strictness)” (*MM*, 6:436).

Because we all start out with tendencies to evil, we are liable to react to our constant humiliations either with despondency, or with the consoling thought that, while we have not lived up to the moral law, at least we have done as well or better in living up to its dictates than so-and-so. Once we start down this path, though, we end up as adequacy fantasists who judge ourselves on the basis of our relative worth, or who judge our moral worth using others as a standard.

Once an adequacy fantasist, one can quite easily lapse into exceptionalism. If you compare yourself to a group of people, but find yourself starting to lag behind them in moral worth, you can rectify the situation simply by coming up with explanations for why their moral worth is inferior to your own:

> There are now only two ways left of getting even with the other’s perfections. Either I seek to acquire those perfections of his for myself as well, or I try to diminish them. Whether I enlarge my own perfection, or lessen his, I always come out the better man. Now since the latter comes easiest, men will sooner diminish the other’s perfections than enhance their own. This is the origin of jealousy. When men compare themselves with others, and find these perfections, they become jealous of every perfection they perceive in the other, and try to diminish it, so that their own may stand out the more. This is disparaging jealousy. But if I try to add to my perfections, making them equal to the other’s, this is emulating jealousy. (*LE-C*, 27:436-7)

That is, rather than try to keep up with those around her, the adequacy fantasist thinks she remains even with others by discounting their merits.

If the adequacy fantasist pursues this path doggedly, she can get away with repeated lapses while also convincing herself that she is doing fine, or even better than others. Eventually, she will think that her perfect obligations are rather minor in

significance compared to actions that advance her merit, and will give herself outs from performing them:

> Are there such trifles as could be seen as transgressions, but which, on account of their unimportance, would not be accountable, and thus might even be permitted? But there are absolutely none such, though in Jesuit casuistry they are accepted *sub voce* peccadillo (from which we get the word "bagatelle"). For, though, in the individual case, the consequence and effect may certainly be a small evil, the maxim adopted by the agent to perform the action, in his determination by the laws of freedom, still remains a large one, and unlimited in its consequences. It is an established fact, that nobody starts off with the grossest crimes, but has been seduced into them by steps which had their basis in subjective principles. It is a small thing when a child thoughtlessly hits another, but habit implants a lack of sensitivity here, and the offender no longer feels anything. From this come steps to acts of violence, and with other maxims concurring in the process, the child can become a murderer. (*LE-V*, 27:557)

6.3. Conclusion

As we have seen, there are a variety of types of evil personality, as many as there can be different fundamental standards of valuation. One can view status as the most important determinant of relative worth; one can measure moral worth by comparing oneself to others rather than to the moral law; a person can think that inner worth comes in degrees, and correspondingly brings with it different obligations. And there are more standards we have not explored – a person could measure inner worth by relative worth; or she may measure her moral worth by comparing herself to the moral law, but conclude that people who are morally worthier than others also have more inner worth as well; and so on.

Whatever standard a person has, though, the reason she has it is that she has an evil *Gesinnung*. In every case, a person's evil *Gesinnung* is the decision to subordinate her Moral Maxim to the Prudential Maxim. This makes her judge satisfying at least some sensible desires as better than discharging some of her moral obligations, and acting on this judgment will set into motion processes that encourage her to construct a

rationalizing scaffolding around this decision and others like it, which, when complete, amount to an evil self-conception of some kind or another.

Finally, this evil self-conception makes certain ways of acting normative, such that acting in those ways will seem to promise her the most pleasure, even if, as a matter of fact, more sensible pleasure could be had by doing something different. This allows for the performance of evil actions that are apparently selfless. Kant's theory of evil can thus accommodate the objection that it makes evil action overly simple, just as it can deal with many of the other standard objections to it.

---

[1] "There is no one – not even the most hardened scoundrel, if only he is otherwise accustomed to use reason – who, when one sets before him examples of honesty of purpose, of steadfastness in following good maxims, of sympathy and general benevolence … does not wish that he might also be so disposed" (*G*, 4:454).
[2] Conscience "is practical reason holding the human being's duty before him for his acquittal or condemnation in every case that comes under a law" (*MM*, 6:400).
[3] "The difference between the correct and the errant conscience lies in this, that error of conscience takes two forms, {error of fact} and {error of law}. He who acts according to an errant conscience is acting conscientiously and if he does so, his action may be defective, but cannot be imputed to him as a crime" (*LE-C*, 27:354-55).
[4] See *G*, 4:436; *LE-C*, 27:407; and *MM*, 6:434-35.
[5] See *G*, 4:436 and *MM*, 6:436.
[6] See *Rel*, 6:27-28.
[7] See, e.g., *G*, 4:440.
[8] See *MM*, 6:435.
[9] See also *CPrR*, 5:157, 161; *LE-C*, 349; *LE-V*, 27:609, 610, 621, 622, and 675; and *MM*, 6:435.
[10] Kant writes in the *Critique of Practical Reason* that "certainty of a disposition in accord with this law is the first condition of any worth of a person" (*CPrR*, 5:73). Admittedly, it is not clear that Kant had yet developed the idea of good and evil *Gesinnungen* at this point; however, certain passages from the second *Critique* (viz., *CPrR*, 5:140 and 143) lead me to believe that he had come rather close.
[11] See *Rel*, 6:61, 71, and 76-77.
[12] See, e.g., *G*,4:401, 406, and 407.
[13] See *G*, 4:406 and Johnson 1996, 326: "Kant's view is simply that actions either are or are not of moral worth, without gradation."
[14] See *LE-C*, 27:359; *CPrR*, 5:84-85; and *MM*, 6:458-59.
[15] See, e.g., *LE-V*, 27:695 and *MM*, 6:458-59.
[16] Presumably, "understanding" in this sense refers to the faculty that encompasses reason, judgment, and the understanding (more narrowly conceived), not just the understanding in the narrow sense. See *Ant*, 7:196-97.
[17] See *G*, 4:429.
[18] Actually, I take every action done from respect to have equal moral worth (an action either has moral worth or not); an action will draw more or less respect from us, though, depending on its level of *merit*, which I address in section 3.4.2 of this chapter.

[19] See, for example, *G*, 4:493 ("It is impossible to think of anything at all in the world, or indeed even beyond it, that could be considered good without limitation except a **good will**") and *LE-C*, 27:344 ("Freedom is … the inner worth of the world").
[20] "To satisfy the categorical command of morality is within everyone's power at all times" (*CPrR*, 5:36-37).
[21] Before "a humble common man in whom I perceive uprightness of character in a higher degree than I am aware of in myself *my spirit bows*, whether I want it or whether I do not" (*CPrR*, 5:76-77).
[22] I take a perfection simply to be a trait that can be regarded as good for some purpose: "as a concept belonging to *teleology*, {the concept of perfection} is taken to mean the harmony of a thing's properties with an *end*" (*MM*, 6:386). Thus, a person's intelligence is a perfection, as is her wit, her strength, her appearance, etc. It is not clear whether a person's humanity can count as a perfection, because it is not clearly a trait (though of course it is useful for attaining to some ends); however, I see no reason why people cannot treat it as a perfection.
[23] As an example of misdirected appraisal-respect, there is Kant's (otherwise puzzling) remark that:

> A man … so far as he is independent of others and has resources, is an object of respect; for a man loses his worth if he depends on others. It is already natural to respect a person less if he depends on others; but if, in turn, he has others at his command, like an officer, that restores the situation. So a common soldier or a servant is less respected. … This worth which arises from independence is merely negative; the positive worth conferred by wealth stems from the power it gives. (*LE-C*, 27:398-9)

[24] See *G*, 4:421-22.
[25] As Kant puts it:

> We have reason to harbor a low opinion of our person, but in regard to our humanity we should think highly of ourselves … This low opinion of our person arises … from comparison with the moral law, and there we have reason enough to humble ourselves. But in comparison with others, we have no reason to entertain a poor opinion of ourselves, for I can just as well possess worth as anyone else. (*LE-C*, 27:348-49)

[26] "{W}ell-meant strictness in determining genuine moral import in accordance with an uncompromising law … greatly lowers self-conceit in moral matters, and humility is not only taught but felt by anyone when he examines himself strictly" (*CPrR*, 5:154).
[27] In Kant's words:

> Comparison with others in determining our own worth can thus be aimed only at self-instruction concerning our value … It is a duty here, to seek out the good that we can discern in their actions, for the use of it really consists in this, that now their actions become motives to prod us into the practice of virtue, in that we thereby become assured that in comparison with the law, and the fulfillment of it achieved by others, our practical virtue is still weak, or in some degree may surpass others. (*LE-V*, 27:703-4)

[28] See *G*, 4:397.
[29] See also *LM-L*$_1$, 28:247; *LM-M*, 29:890-91 and 894; *CPrR*, 5:9n; and *Ant*, 7:231.
[30] See *LE-C*, 27:257-8.
[31] See *MM*, 6:386-87.
[32] See also *LE-C*, 27:457 and *LE-V*, 27:678-79.
[33] It "is indeed impossible that in the sight of God, as the law of the highest morality, we can do more than is incumbent, since in regard to Him, everything is required; but in relation to other men, we can certainly have merits" (*LE-V*, 27:665).
[34] See Johnson 1996, 313-16.
[35] See Johnson 1996, 316.
[36] See, e.g., *LE-C*, 27:410; *G*, 4:424; *LE-V*, 27:622 and 665; and *MM*, 6:390 and 448.
[37] Presumably, performing a wide duty out of a morally forbidden motivation – such as the desire to trick people into thinking you meant them no harm – would disqualify that action from being meritorious (see Johnson 1996, 320-21 for evidence that this was indeed Kant's view).
[38] See also *LE-V*, 27:622.
[39] See Johnson 1996, 315.
[40] See also *LE-V*, 27:664-65.
[41] See also *LE-V*, 27:666.

[42] What if the reason it is difficult for a person to perform an action is that she has lots of evil motivations that make it hard for her even to discharge a narrow duty? In such a case, she would not deserve praise just for discharging what she owed, but if she did her duty out of respect, or even out of a benign motivation, such as love, then she would act meritoriously.
[43] This is an unfortunate choice of word, given how many other ways in which Kant uses "respect". What Kant means, though, by associating respect with the minimum of praise we can give someone, is the contradictory of contempt. That is, we respect someone, in this sense, when we do not contemn her. For passages where Kant uses respect in this way, see *LE-C*, 27:409 and 410; and *MM*, 6:448-49.
[44] See *CPrR*, 5:155.
[45] "It is easy to show that the human being is actually guilty of many **inner** lies, but it seems more difficult to explain how they are possible; for a lie requires a second person whom one intends to deceive, whereas to deceive oneself on purpose seems to contain a contradiction" (*MM*, 6:430).
[46] See *Rel*, 6:41-42.
[47] Remember, though, a solipsist could subordinate her own desire for happiness to respecting her perfect duties to others.
[48] See *MM*, 6:392-93, where Kant says acting on the maxim of fulfilling your duties out of respect for the moral law is a wide duty, and also part of your perfecting yourself. However, since part of fulfilling your duties is fulfilling your imperfect duties to others, someone who neglects those will neglect this as well.
[49] Note that a solipsist who accepts such a principle will think that she has imperfect duties of beneficence to people whom she judges to have relative worth as high or higher than she. However, she might (or might not) be blinded by her self-love from recognizing the equal (or higher) level of relative that others may have, allowing her to happily cater to her own desires.
[50] Of course, a person might adhere to a principle that indeed commits her to violating others' perfect duties in some unusual situations, and might not know this because she has not yet encountered such situations. Such a person would *act* solipsistically, but would not in fact be a solipsist.
[51] Thus, I take the imperfect duties to promote others' happiness and develop one's own perfections to be duties to *maximize* both these things. While it is true that Kant writes that morality "leaves a playroom (*latitudo*) for free choice in following (complying with) the law", he clarifies that "a wide duty is not to be taken as permission to make exceptions to the maxim of actions but only as permission to limit one maxim of duty by another (e.g., love of one's neighbor in general by love of one's parents), by which in fact the field for the practice of virtue is widened" (*MM*, 6:390).
[52] See *LE-V*, 27:703-4: "Comparison with others in determining our own worth can thus be aimed only at self-instruction concerning our value … It is a duty here, to seek out the good that we can discern in their actions, for the use of it really consists in this, that now their actions become motives to prod us into the practice of virtue".
[53] See *CPrR*, 5:154.
[54] Kant writes that the "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*" (*CPrR*, 5:74). In other words, self-love (i.e., moral self-love, or *philautia*) is the propensity to self-conceit.
[55] Indeed, it seems *impossible* for it to be true, for anyone who uses this false standard of evaluation instead of the moral law shows herself to have no moral worth (although she could still perform actions that have moral worth).
[56] The sentence right after this equates merit to moral perfections, suggesting that the merit Kant has in mind here is moral merit.
[57] See also *LE-V*, 27:610-11.
[58] See *LE-V*, 27:575: "a conscience consists in the ability to impute one's own *factum* to oneself, through the law itself".
[59] Kant mentions the examining conscience (or *conscientia antecedens*, or "warning conscience") at *MM*, 6:440; *LE-H*, 27:43-44; *LE-C*, 27:356; and *LE-V*, 27:616-18. He mentions the accompanying conscience (or *conscientia concomitans* [conscience that monitors the act]) at *LE-H*, 27:43 and *LE-C*, 27:356. He mentions the judging conscience (or *conscientia consequens*) at *LE-H*, 43-44; *LE-C*, 356; and *LE-V*, 27:616, 617-18.

It should be noted that in the Vigilantius lecture notes, Kant goes on to deny the existence of the accompanying conscience:

> Professor Kant says that we cannot properly conceive the possibility of a *conscientia concomitans*, if it is to operate during the action. For the latter, as consequence, always presupposes already an approbation on the part of conscience, and we are merely postulating in our actions two stages that follow directly on each other, and therefore call them *concomitantes*. (*LE-V*, 27:617)

Kant seems to think that there are not three stages—one's mental state before the action, one's mental state during the action, and one's mental state after the action—but rather just two, a mental state before, and one after, the action. The idea seems to be that there is no period of time that exists during the commission of an action; even if an action—say, sawing through a tree—takes a long time to complete, what is assessed by conscience is a person's thinking before and after the moment of deciding to act. The actual moment of deciding to act is only a moment, and so not one accompanied by any thinking.

I do not agree with Kant here, because I think it is useful (for attributions of responsibility) to posit an accompanying conscience that pleads with someone not to do (or to continue) what she is doing; for example, the longer someone persists in an activity she knows to be wrong, the worse her action is.

[60] There are two ways in which a person might need to learn conscientiousness. First, a person must become attentive to the voice of conscience in the first place, i.e., she must learn as a child what that little voice in her head is, and train herself to listen to it (this is evidenced by the fact that Kant thinks of conscience as a moral predisposition, i.e., a capacity that needs first to be activated and then cultivated (*MM*, 6:438). Second, if a person practices the same vice over and over, her conscience will become inured to it and will cease trying to warn her off from it (more on this in the next paragraph); if this happens, she needs to turn her unconscientiousness in relation to this vice into conscientiousness.

[61] The "greater the villain, the less does conscience plague him, for a tormenting conscience is still the remnant of a good disposition" (*LE-M*, 29:623).

[62] See *CPrR*, 5:158-59.

Works Cited

Adams, Robert M. (1987). "Must God Create the Best?" in *The Virtue of Faith: And Other Essays in Philosophical Theology* (Oxford: Oxford University Press), 51-64.

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

——— (1996). "Reflections on the Banality of (Radical) Evil: A Kantian Analysis", in *Idealism and Freedom: Essays on Kant's Theoretical and Practical Philosophy* (Cambridge: Cambridge University Press), 169-182.

——— (2001). "Ethics, Evil, and Anthropology in Kant: Remarks on Allen Wood's *Kant's Ethical Thought*", *Ethics*, Vol. 111, No. 3 (April: 594-613).

——— (2002). "On the Very Idea of a Propensity to Evil", *The Journal of Value Inquiry*, Vol. 36, Nos. 2 and 3 (June: 337-48).

Anderson-Gold, Sharon (1991). "God and Community: An Inquiry into the ReligiousImplications of the Highest Good", in Philip J. Rossi and Michael J. Wreen, editors, *Kant's Philosophy of Religion Reconsidered* (Bloomington, IN: Indiana University Press), 113-31.

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

Barth, Karl (1969). *Protestant Thought from Rousseau to Ritschl*, translated by Brian Cozens (New York: Simon and Schuster).

Beck, Lewis White (1960). *A Commentary on Kant's* Critique of Practical Reason (Chicago, IL: The University of Chicago Press).

Bernstein, Richard J. (2002). *Radical Evil: A Philosophical Interrogation* (Malden, MA: Polity Press).

Brewer, Talbot (2002). "Maxims and Virtues", *The Philosophical Review*, Vol. 111, No. 4 (October: 539-72).

Bubner, Rüdiger (2001). "Another Look at Maxims", in Predrag Cicovacki, ed., *Kant's Legacy: Essays in Honor of Lewis White Beck* (Rochester, NY: University of Rochester Press).

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

Caswell, Matthew (2006a) "Kant's Conception of the Highest Good, the *Gesinnung*, and the Theory of Radical Evil", *Kant-Studien*, Vol. 97, No. 2 (June: 184-209).

——— (2006b). "The Value of Humanity and Kant's Conception of Evil." *Journal of the History of Philosophy*, Vol. 44, No. 4 (October: 635-663).

Crocker, Jennifer, Shawna J. Lee, and Lora E. Park (2004). "The Pursuit of Self-Esteem: Implications for Good and Evil" in Arthur G. Miller, editor, *The Social Psychology of Good and Evil* (New York: The Guilford Press), 271-302.

Darwall, Stephen L. (1977). "Two Kinds of Respect", *Ethics*, Vol. 88, No. 1 (October: 36-49).

Deigh, John (1983). "Shame and Self-Esteem: A Critique", *Ethics*, Vol. 93, No. 2(January: 225-45).

Departments of Health and Human Services: Centers for Disease Control and Prevention. *Alcohol: Frequently Asked Questions* URL = <http://www.cdc.gov/alcohol/faqs.htm#10>. Accessed on June 29, 2007.

Engstrom, Stephen (1988). "Conditioned Autonomy", *Philosophy and Phenomenological Research*, Vol. 48, No. 3 (March: 435-53).

Fackenheim, Emil L. (1954). "Kant and Radical Evil", *University of Toronto Quarterly*, Vol. 23, No. 1 (Winter: 339-353).

Fara, Michael (2006). "Dispositions", in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition)* URL = <http://plato.stanford.edu/archives/fall2006/entries/dispositions/>.

Frierson, Patrick R. (2003). *Freedom and Anthropology in Kant's Moral Philosophy* (Cambridge: Cambridge University Press).

——— (2005). "Kant's Empirical Account of Human Action", *Philosophers' Imprint*, Vol. 5, No. 7 (December: 1-34).

——— (2006). "Character and Evil in Kant's Moral Anthropology", *Journal of the History of Philosophy*, Vol. 44, No. 4 (October: 623-34).

Grimm, Stephen R. (2002). "Kant's Argument for Radical Evil". *European Journal of Philosophy*, Vol. 10, No. 2 (August: 160-77).

Hare, John E. (1996). *The Moral Gap: Kantian Ethics, Human Limits, and God's Assistance* (Oxford: Clarendon Press).

Herman, Barbara (2007). "Rethinking Kant's Hedonism", in *Moral Literacy* (Cambridge, MA: Harvard University Press), 176-202.

Jansen, Ludger (2007). “On Ascribing Dispositions”, from Max Kistler, Bruno Gnassounou (eds.), *Dispositions and Causal Powers* (Aldershot: Ashgate, 2007), 161-77.

Johnson, Robert N. (1996). “Kant’s Conception of Merit”, *Pacific Philosophical Quarterly*, Vol. 77, No. 4 (December: 310-34).

——— (2007). “Value and Autonomy in Kantian Ethics”, forthcoming in Russ Shafer-Landau, ed., *Oxford Studies in Metaethics: Volume II* (Oxford:Oxford University Press). URL = <http://web.missouri.edu/~johnsonrn/osme.pdf>.

Kitcher, Patricia (2003). “What is a Maxim?”, *Philosophical Topics*, Vol. 31, Nos. 1 and 2 (Spring and Fall: 215-43).

Korsgaard, Christine M. (1996). “Morality as Freedom”, in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press), 159-87.

Kuehn, Manfred (2001). *Kant: A Biography* (Cambridge: Cambridge University Press).

Lewis, C. S. (1996). *Mere Christianity: Comprising* The Case for Christianity, Christian Behaviour*, and* Beyond Personality (New York: Touchstone).

Louden, Robert B (2000). *Kant’s Impure Ethics: From Rational Beings to Human Beings* (Oxford: Oxford University Press).

McCarty, Richard (2006). “Maxims in Kant’s Practical Philosophy”, *Journal of the History of Philosophy*, Vol. 44, No. 1 (January: 65-83).

Michalson, Jr., Gordon E. (1990). *Fallen Freedom: Kant on Radical Evil and Moral Regeneration* (Cambridge: Cambridge University Press).

Morgan, Seiriol (2005). “The Missing Formal Proof of Humanity’s Radical Evil in Kant’s *Religion*”, *The Philosophical Review*, Vol. 114, No. 1 (January: 63-114).

Pereboom, Derk (1996). "Kant on God, Evil, and Teleology", *Faith and Philosophy*, Vol. 13, No. 4 (October: 508-33).

——— (2006a). "Source Incompatibilism and Alternative Possibilities", unpublished manuscript. Available at URL =<http://www.calvin.edu/academic/philosophy/virtual_library/articles/pereboom_derk/source_incompatibilism_and_alternative_possibilities.pdf>.

——— (2006b). "Kant on Transcendental Freedom", *Philosophy and Phenomenological Research*, Vol. 73, No. 3 (November: 537-67).

Reath, Andrews (2006a). "Kant's Theory of Moral Sensibility: Respect for the Moral Law and the Influence of Inclination", in *Agency and Autonomy in Kant's Moral Theory: Selected Essays* (Oxford: Clarendon Press), 8-32.

——— (2006b). "Hedonism, Heteronomy, and Kant's Principle of Happiness", in *Agency and Autonomy in Kant's Moral Theory: Selected Essays* (Oxford: Clarendon Press), 33-66.

Scanlon, T. M. (1998). *What We Owe to Each Other* (Cambridge, MA: The Belknap Press of Harvard University Press).

Schroeder, Mark (2005). "The Hypothetical Imperative?", *Australasian Journal of Philosophy*, Vol. 83, No. 3 (September: 357-72).

Singer, Peter (1995). *How Are We to Live?: Ethics in an Age of Self-Interest* (Amherst, MA: Prometheus Books).

Stark, Werner (2003). "Historical Notes and Interpretive Questions about Kant's Lectures on Anthropology", translated by Patrick Kain, in Brian Jacobs and Patrick Kain,

editors,*Essays on Kant's Anthropology* (Cambridge: Cambridge University Press), 15-37.

Sussman, David (2005). "Perversity of the Heart", *The Philosophical Review*, Vol. 114, No. 2 (April: 153-77).

Unger, Peter (1996). *Living High and Letting Die: Our Illusion of Innocence* (Oxford: Oxford University Press).

Van Inwagen, Peter (2002). *Metaphysics: Second Edition* (Boulder, CO: Westview Press).

——— (2006). *The Problem of Evil: The Gifford Lectures Delivered in the University of St. Andrews in 2003* (Oxford: Oxford University Press).

Wood, Allen W. (1996). "Self-Love, Self-Benevolence, and Self-Conceit", in Stephen Engstrom and Jennifer Whiting, editors, *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty* (Cambridge: Cambridge University Press), 141-61.

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

——— (2005). *Kant* (Malden, MA: Blackwell Publishing).