A PROLEGOMENON ON EVIL:

"WHAT DOES IT MEAN TO BE EVIL?"

By

MARK SMITH FERGUSON

Bachelor of Arts in Philosophy

University of Central Oklahoma

Edmond, Oklahoma

2009

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF ARTS
May, 2013

A PROLEGOMENON ON EVIL:

"WHAT DOES IT MEAN TO BE EVIL?"

Thesis Approved:

Dr. Lawrence Pasternack

Thesis Adviser

Dr. Rebecca Bensen-Cain

Dr. Eric Reitan

Name: Mark Ferguson

Date of Degree: MAY 2013

Title of Study: A PROLEGOMENON ON EVIL: "WHAT DOES IT MEAN TO BE EVIL?"

Major Field: Philosophy

ABSTRACT: This thesis is an open-ended inquiry exploring the thought processes within evil actions as it relates to agent judgment and motivation. Largely theoretical in nature, the goal is to better understand the inner workings of evil agency. It is not the purpose of this thesis to ascertain or support a normative ethical theory of evil but rather investigate through metaethics, moral psychology, and ultimately Kantian ethical theory, how evil surfaces in action. That being said, the question which occupies this thesis is "What does it mean to be evil?" Everyone is familiar with the term "evil," but the notion has many connotations in moral discourse. Chapter one establishes a working definition of evil by considering the ways in which people are generally motivated to act. Evil is conceptualized into two distinct categories: perverse and pure evil. This distinction incites considerable debate—especially the latter conceptualization. Whether purely evil motivations are possible or conceptually coherent will serve to dominate a large part of this chapter and the rest of this thesis. Chapter two supplies a metaethical context to evaluating evil motivations in agents—motivation internalism and externalism. These metaethical positions explore whether moral motivations are fundamentally inherent to one's expressed judgments. In other words, is it possible that moral judgments can fail to motivate someone to act? This added dimension, though, only seems to heighten the controversy because pure evil involves principally choosing to do evil for itself. Motivation internalism seems at odds with certain motivational structures, especially the purely evil agent. By highlighting the conflict between internalism and externalism, the subtleties of agent motivation and judgment lead to a more nuanced account of evil. Chapter three introduces Immanuel Kant's account of evil in *Religion within the Bounds of Bare Reason* and how it may provide a possible solution to the troubles of motivation internalism. Kant's three grades of evil and subsequent views on diabolism are susceptible to an interpretation that frames pure evil as a quasi-diabolism in which the moral law is motivationally inverted. This solution attempts to expand Kant's account while preserving his fundamental a priori principles.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

CHAPTER I

A Conceptual and Psychological Analysis of Evil

There is something about evil that fascinates the human mind. From fictional novels, motion pictures, and popular role-playing games like *Dungeons and Dragons*, an extraordinary amount of interest has been focused on the moral intricacies of being evil. Even so, its fascination in literature and pop culture is only further beset by the profundity of the concept philosophically. Evil is far from the ordinary in thought and action due to its metaphysical (or ontological) status. Yet, within ordinary moral discourse, it is a concept that has been utilized extensively in moral assessment as a sign of condemnation.[1] It is the task of philosophers in the field of metaethics (with perhaps the assistance of other fields such as moral psychology) to look beneath moral evaluations and discover what it really means to be evil.

Discussions about evil tend to be dominated by normative concerns—whom or what actions ought to be regarded as evil according to particular moral standards. These accounts, though, largely fail to penetrate the depths of an agent's thought processes—the development of one's beliefs, judgments, and motivations which leads to an "evil act". There is a tendency by laypeople to hastily ascribe specific qualities to evil, especially those with religious or theological overtones.[2] Furthermore, the layperson often focuses on action itself at the cost of ignoring the

[1] Generally speaking in moral discourse, the term "evil" designates actions which "bad" by itself cannot adequately cover. Thus, the statement "X is evil" can be taken as simply a stronger and more economical (as well as emotionally-loaded) expression of the statement "X is really, really, really bad."

[2] For the purposes of staying on topic, the term "evil" will not be constrained by a fixed account. At the very least, some actions can be called evil while others less so or not at all. Furthermore, normative

motivations and thought process that lead to the action. As such, the term "evil" is often invoked without a second thought to any deeper meaning. Underlying the common use of the term and its normative meaning designated by various ethical theories is a significant conceptual and psychological groundwork—the "how" and "why" of evil. These issues add to a larger metaethical picture. This thesis will explore the motivational structure of evil agency and whether there is a different way people can think about evil.

This chapter will clarify the term "evil" and its significance in moral deliberation. In addition, section 1.1 will outline and defend two main conceptualizations (or classifications) of evil in moral discourse: perverse and pure evil. Perverse evil signifies the common, generalized, usage of evil and is intended to assign negative conditions to agents. Pure evil represents a more technical and specified kind of evil that must be explained (and defended) further. This conceptual distinction has direct implications on developing a moral psychology of evil and impacts the metaethical theories discussed in later chapters.

Section 1.2 explores a position notably represented by Elizabeth Anscombe and David McNaughton that couches evil agency in negative terms, as embittered and maliciously rebellious closet lovers of the good. Milton's Satan, in *Paradise Lost*, will be used as a case study to ascertain the motives and psychological character of perverse and purely evil agency. This is done in the hopes of persuading the reader of the view above that there are at least two different ways of thinking about evil. Nonetheless, this chapter is but a precursor to a larger issue. These conceptual and psychological inquiries are critically influenced by metaethical theory, in

---

disputes about whether evil is dependent upon a religious meaning or can be just as intelligible in a secular context will be put aside.

particular the debate between motivation internalism and motivation externalism. Sections 1.3 and 1.4 establish this context and certain benchmarks are formulated to judge the impact both metaethical views have on understanding the nature of evil. Starting with this chapter, evil will be examined and stripped beneath the surface to its core.

1.1 The Multifarious Parameters of Evil

Evil has many faces and can take form in unexpected ways. If Plato's rendition of Gyges' Ring is indicative of human tendencies towards wrongdoing, then perhaps the prospect of evil rests within every single person. Erich Fromm points out that one should be careful of assuming that evil must look its part. He warns that "as long as one believes that the evil man wears horns, one will not discover an evil man."[3] Any investigation into the inner workings of evil must heed this advice and avoid being a slave to appearances. Even an apple that is rotten to the core will still remain in appearance, for some time, an ordinary apple. The appearance or visible signs of evil in a person can only be interpreted to a certain extent if granted further information. While others may take a more empirical route to understanding evil, the approach of this thesis will be decidedly conceptual and intermixed with moral psychology. Grasping the nature of evil is subject to many interpretations, but examining the issue in terms of the parameters of judgment and motivation that precede action might produce better results. To grasp evil in this way, one must look within moral discourse and see the possible manifestations of evil among moral agents.

From a wholly conceptual standpoint, evil can be divided into two complementary, yet distinctive, categories: perverse (or instrumental) evil and pure evil. This division is by no means original.[4] Perverse evil, as suggested by Irit Samet-Porat, is the performance of evil acts under the guise of moral certitude or permittance.[5] The perversely evil agent can embody a wide range of mindsets. There are three ways that immediately come to mind. The individual can have a twisted moral code that rationalizes the act to be morally justified, have an alternative motive to commit evil acts (such as the desired utility of the act), or have an emotional breakdown of some

[3] Erich Fromm, *The Anatomy of Human Destructiveness* (Greenwich: Fawcett Crest, 1973), 480.
[4] See the following works: Colin McGinn, *Ethics, Evil, and Fiction* (Oxford: Clarendon Press, 1997); Daniel M. Haybron, "Evil Characters," *American Philosophical Quarterly* 36, no. 2 (1999); and Irit Samet-Porat, "Satanic Motivations," *The Journal of Value Inquiry,* no 41(2007).
[5] Samet-Porat, "Satanic Motivations," 78.

kind that dramatically deforms their moral character. Perverse evil can arise in these different ways but the conceptual thought process behind it remains the same. Haybron similarly uses the term "corrupt evil" to describe moral agents that culpably "choose evil" when they could have done the opposite, perhaps due to circumstances or an otherwise obscured character defect.[6] Evil in all these senses is performed instrumentally for some reason other than its own sake.

Pure evil, on the other hand, is the performance of evil acts for their own sake—mirroring opposite the old adage "be good for goodness sake". The purely evil agent operates on the same level intellectually as any other normal person in that he/she can make the same moral evaluations, but is motivated to perform evil as evil rather than good as good. In other words, these agents are attracted to evil for itself. They, unlike perversely evil agents, cannot be straightforwardly accused of ignorance or some form of weakness.[7] Purely evil agents care about morality, just not the side of the moral issue others would expect. Satan is often mentioned as a paradigm case of doing evil on principle; but this case is not without some controversy.[8]

Most, if not all, instances of evil in ordinary matters are some derivable form of perverse evil. Even the horrific actions of Hitler and his associates are representative of perverse or instrumental evil. They did not commit genocide purely for the sadistic desire to kill others. Hitler not only had an agenda which included the desire of creating a master race, but also acted in such a way that he believed his actions were justified as good. People who have done far lesser offenses than genocide have adopted the very same mindset, like the common criminal or murderer. The major difference between such people and Hitler is in degree, but not in kind. Conceptually, both of their actions are a means to some desired end not resulting from pure principle. Hitler presumably considered the wholesale destruction of an entire people the best

[6]Haybron, "Evil Characters," 141.
[7]Samet-Poret, "Satanic Motivations," 79-80.
[8] Section 1.2 will be devoted, in part, to this very issue.

way to achieve his end goals. Further, it would be strange to think—at least in the case involving Hitler—that someone would murder or commit genocide just for the sake of doing it.

Perhaps the reason why most evil can be explained in terms of perversion (or corruption) is because of the nature of the agent's judgments themselves. The fact that judgments can go awry at every possible step in the deliberation process suggests that an agent's motivation to commit murder or orchestrate genocide is very much a possibility, though hopefully a rare one. Hitler's genocidal agenda was spurred by his nationalism and an ideology bent on racial purity (i.e. ethnocentrism). Despite his role in the deaths of millions, Hitler seemed to have no overwhelming environmental cues from childhood that set him apart from anyone else; he had a warm and loving mother (though his father was a strict authoritarian).[9] The question then is what caused him to show such disregard for others' lives?

Fromm suggests that Hitler's friendly and amicable nature was a role which he likely valued for its usefulness.[10] Unlike the Dr. Jekyll and Mr. Hyde characters from Robert Louis Stevenson's 1886 short story, Hitler could control which "personality" suited his situation. But the assumption should not be made that *all* of his kindness was mere veneer. Though Hitler possessed little to no affection for other people, he had a rather surprising care for animals—especially his dog.[11] In any case, with perversely evil agents like Hitler, subscribing to an abusive ideology or belief-system can change how one's values are fulfilled. When people succumb to bigotry and adopt a destructive agenda, such agents act out of misguided self-righteousness with "good" being used to identify the fulfillment of their (perverse) projects.

It is important to note that kindness does not entail the work of a benevolent agent as much as cruelty does not entail the work of an evil agent. Hitler's actions, cruel and destructive

[9] Fromm, *Anatomy of Human Destructiveness*, 413-422.
[10] Ibid., 470.
[11] Ibid., 454.

in practice, had their beginnings in thought and were roused by motivation. One's actions after all do not exist in a vacuum. They are traceable to an agent's thought process and collection of motivations that weighs some actions over others. The evaluation of evil acts as presented to the naked, empirical eye is incomplete. Such an analysis taken at face value would fail to heed the advice mentioned earlier and would only scratch the surface of evil agency's inner workings. This also seems to relate to some psychological structure underlying evil, within both perverse and pure forms, that will require further investigation in subsequent sections.

There are other instances of perverse evil with more complex parameters that need to be mentioned. In particular, the anti-villain model[12] offers some insight into complicated, evil personas. Anakin Skywalker's character development in the *Star Wars* saga is a peculiar example of the ambiguity that can exist between good and evil characterizations. As a child, it was believed that Anakin was the "chosen one" of the Jedi order who would bring balance to the force. But a series of unfortunate events, as well as some exposed character flaws and extremely poor decisions, changed Anakin and his future.

Slavoj Zizek, in *The Parallax View*, argues that Anakin's turn to evil was (or rather should be viewed as) due to his excessive attachment to the good.[13] Anakin's fall began with his ambition to excel in the Jedi ranks and then heightened with his desire to save Padme (from his dreams about death during childbirth). This resulted in Anakin's need for power (at any cost) to control their destinies.[14] That is, he turned to evil through "an overwhelming desire to intervene,

[12] The anti-villain is an agent that is ambiguously evil. He/she has desired ends that seem evil or nefarious but also possess heroic traits or virtues that make their ends more ethically-oriented than most villains. Like the concept of the anti-hero, these are literary devices intended to blur the lines between the "good guys" and the "bad guys" in story-telling narratives.

[13] Slavoj Zizek, *The Parallax View* (Cambridge: MIT press, 2006), 100-103.

[14] Richard Corliss, "Dark Victory," *Time Magazine*: April 22, 2002.
The following comments by George Lucas in this article seem to complement the assessment above: "He turns into Darth Vader because he gets attached to things. He can't let go of things. He can't let go of his mother; he can't let go of his girlfriend. He can't let go of things."

to do Good, to go to the end for those he loves (Amidala)."[15] These attachments, created out of Anakin's love and devotion, took over and ultimately twisted his character. The moment in which Anakin's name changed to "Darth Vader" signals the culmination of Anakin's moral transformation. Mania, at first a symptom of Anakin's attachment to good (i.e. his love for Padme), became a driving force for perverse evil causing him to overcompensate within his own moral evaluations.

With Emperor Palpatine's manipulation, Anakin became enthralled by the rage and aggression brewing within him as he desperately attempted to prevent his dreams of Padme's death from coming true. These things led to appalling acts of infanticide, the destruction of the Jedi order, and the conclusion of Anakin's fears culminating in a self-fulfilling prophecy. The loss Anakin suffered from his mother's death earlier in *Attack of the Clones* (2002) makes his mania sympathetic and his turn to perverse evil all the more tragic. While people would hardly condone Anakin's actions, such as the slaughter of children at the Jedi temple in *Revenge of the Sith* (2005), the tragic nature of Anakin and similar characters are not beyond sympathy. By a strange twist of fate, Anakin's love for others led him to commit great evils.

The anti-villain model adds a dynamic element to evil in moral discourse and further demonstrates the various kinds of perverse evil. The social, psychological, and interpersonal conditions of evil agency are boundless in terms of literary resources. Many stories and character models, with a few exceptions, mirror or emulate in distinct ways people's lives in the real world. Cinematic and literary characters, though they may be a product of the mind, can possess some realism because the reader or viewer can relate to them like any other human being. They can synthesize the details of others and make a personal connection with those characters.

[15] Ibid., 101.

Oftentimes these details, fictional or not, demonstrate untold variables which can give rise to evil. If evil can come from the unlikeliest of sources where even an otherwise positive emotion such as love is not exempt from being a catalyst, then perhaps good and evil are dualities that overlap from time to time. When imagination and experience combine with the power of narration, the notion of what is evil (and good) becomes slightly more opaque. The parameters of evil, thus, expand and multiply when the depths of the individual's psyche and one's relationships with other people are explored. The categorizations of perverse and pure evil can be further delineated when additional conditions and circumstances are included.

Zizek's reasoning about the transformation of Anakin to Darth Vader could also be read as criticism of the laymen tendency to speak in absolutist terms. Moral discourse is rife with instances of unconditional statements about good and evil while not carefully recognizing the meticulous processes that lead to making moral decisions. While anyone familiar with ethical theory or who has taken a general philosophy course in ethics will already be aware of the pitfalls of absolutist ethics, it is important to emphasize this point before continuing.

Individual characters hardly fit into ideologically pure models of good and evil, but the ethical construction of the Stars Wars universe seems to support this simplistic dichotomy. With the Jedi and the Sith, George Lucas presents an ethical landscape in which the light and dark sides of the Force are unambiguously distinct and incapable of immersion. Both "sides" are strict in their principles; the Sith seek power and domination, whereas the Jedi wish for peace and harmony. Both also turn to the extreme in the defense of their ideals. The Sith celebrate emotions as a source of strength even to the point of being consumed by it, while the Jedi reject all avenues of sensuality as a dangerous slippery slope and cultivate an ascetic lifestyle. The multifarious parameters of evil outlined in this section hopefully show the error of such rigorous characterizations.

It does not seem farfetched to suggest that the concept of pure evil mentioned earlier arises from this excess in extremes. Although Anakin's lust for power was for the sake of some other end (i.e. love, glory, etc.), one could imagine a different set of parameters that results in a person motivated to act badly just for the sake of doing it. Such a person would surely be far from the ordinary since people presumably want power, love, and glory for some other end (i.e. happiness) in much the same way many people desire money for the expressed purposes of spending it. But, supposing the concept of pure evil is illusory, all evil conduct may be just a matter of perpetuating actions out of some (misguided) selfish endeavor. In that case all evil would be reducible to perverse evil. But that suggestion, at this time, is premature.

In the case of money, is it possible for someone to just want money for its own sake? While currency by itself is nothing but the materials it is made of, its essential value need not be completely tied to its bartering value either. The public value of currency can be co-opted into an intrinsic value on the same level as someone may value health or joy for its own sake. The Ebenezer Scrooge character seems to best embody an agent with the mental inner-workings of desiring money-qua-money. This agent, a plutocrat to the extreme, seems to literally define the possession of money as intrinsically valuable. As someone may deem "bachelor" tautological to "unmarried man," Scrooge intimately associates "money" as tautological to "something desirable for itself". Furthermore, Scrooge's miserly lifestyle up to the point of his ethical conversion serves as testament to his valuation of money as an end to itself. Scrooge hoards money, seeks to possess as much of it as he can, and as a result spends as little of it as necessary even at the cost of his own well-being (e.g. when he adamantly refuses to work the heater in order to save money despite being noticeably affected by the temperature in *A Christmas Carol* (1984)).

In other words, Scrooge does not revel in his wealth but subsists as if money had no value other than being possessed and continually accumulated. His motivational framework does not seem to attribute an instrumental value of money such as using it for pleasure or to acquire power

over others. Scrooge simply did not care about others and presumably only had a care when money was at stake. Perhaps, as a businessman, Scrooge did initially place an instrumental value on money. At some point, though, it ceased to solely have that value for him. But one can only know so much about a person. Even a literary character has limitations to what one can legitimately know about his/her thought processes—which makes this inquiry extremely difficult to discuss.

But this in part demonstrates the importance of exploring in more detail the inner workings of an agent's judgments and motivations as they relate to action so that one can better understand people like Scrooge—as well as how one's own motivations work to produce judgment and encourage certain actions. Scrooge and his penny-pinching ways may seem largely trivial for most people but the prospect of doing evil for evil's sake operates along a similar thought process. The question then arises: Can a similar case also be made for agents that desire evil-qua-evil (i.e. purely evil motivation)? Does pure evil have some test case distinct from perverse evil?

Instances of pure evil, though, are harder to isolate from garden-variety perverse evil. It has been suggested that some types of serial killers or psychopathic murderers—like "thrill killer"[16] Robert Alton Harris—may qualify as realistic models of pure evil.[17] But one could argue that the existence of a harsh childhood upbringing may have shaped them into such monsters. For evil to be done for its own sake, there can be no pretense to instrumental ends underlying the agent's motivations or possible disparity in their moral understanding. Pure evil, to be clearly distinguishable from perverse characterization, must present a clear and conscious desire to do

[16] The term "thrill killing" is premeditated murder, oftentimes of a complete random stranger, motivated primarily by the pure excitement of the act itself with no clear indication of mental instability. One possible example, but nonetheless not a clear indication of the matter, is the famous Leopold and Loeb case in 1924.

[17] Haybron, "Evil Characters," 139.

evil. While there can be secondary or auxiliary reasons for an action, the purely evil agent's primary motivation must be with respect to the knowledge that he/she does X *because* X is evil.

This conceptualization of evil treads into enigmatic waters. The kind of agent that simply commits an evil act, say murder, for its own sake seems beyond comprehension. There is a tendency to suggest that there must be some reason other than pure malevolency that lurks behind the agent's thoughts—that evil emanates from some sort of self-interested code or perverse gratification and thus, cannot be done for its own sake. Perhaps (as suggested above with the Scrooge character) this is due to the difficulty of fully determining the motivations and/or intentions of people from the outside looking in. If, in response to the question of why commit murder, someone were to say, "I just wanted to murder someone today. I had nothing instrumental to gain or lose from such an action. In fact, I knew that it was a despicably evil thing to do," then it would not be unreasonable to be skeptical of that person's own reasons and perhaps sanity.

There might be grounds to question the psychological health of purely evil agents. If many serial killers come from a background of abuse and violent upbringings, then there may not have been sufficient development of a general sense of empathy for others. Charles Darwin, in *The Descent of Man*, argues that the endowment of well-marked social instincts— among them the parental and filial affections—constitute the building blocks of a moral sense or conscience.[18] Feelings of sympathy, friendship, and love arise within social groups in the form of advantageous traits for natural selection. If individuals within a group are amicable toward each other in these ways, then the group as a whole has a better chance of survival and the individuals within to propagate future generations. Sociability, for Darwin, signifies the beginnings of a moral point of view because an individual's behavior includes considering the welfare of others outside of

[18] Charles Darwin, "Origin of the Moral Sense" or "Comparison of the Mental Powers of Man and the Lower Animals", *Decent of Man* (New York: Random House Inc., 1936), (Ch. IV) 471-495.

familial ties, not just for oneself and one's progeny. Serial killers in this respect are sociopathic, for they have failed to reach those points of development and perhaps cannot sincerely appreciate the communication of moral imperatives.

To commit evil acts just for the sake of doing them seems to exemplify a lack of empathy for others which significantly distinguish ordinary agents from the amoral, the perverse, and especially the purely evil. But sociopathy and the qualities it entails are not required—nor do they seem sufficient—to be included in the domain of pure evil. Perhaps these associations are just incidental for some instances of pure evil. As such, until adequate reasoning is presented to the contrary, one should leave sociability and psychological instability as possible but not necessary characteristics of purely evil agency. Like Glaucon's model of the perfectly unjust man in the *Republic*,[19] the purely evil agent should be allowed the same access to the rewards or goods of any other normally-situated person. One should "subtract nothing" from the agent in order to see in full view "his own way of life."[20] Otherwise, one could end up making ad hoc appeals to undermine pure evil and reduce it to a sophisticated, self-delusional variant of perverse evil. Just as Glaucon implored Socrates to avoid detracting from the unjust person the means of power, wealth, friends, and even the skill of a "clever craftsman" so that one can come to terms with injustice itself,[21] purely evil agency should also not be diluted until one has grasped the full extent of its impact in moral discourse and whether it fits within the ethical landscape. Then—and only then—should one tamper with it in light of some objectionable grounds.

Nevertheless, the characteristics that pertain to thrill killers may have some relation to purely evil agency. Gary Watson notes that while thrill killer Robert Alton Harris' abusive childhood does make people less inclined to use some "reactive epithets", it should not exempt

[19] Plato, *Republic*, trans. G.M.A. Grube (Indianapolis: Hackett, 1992), 36 [360e-361b].
[20] Ibid.
[21] Ibid.

Harris from judgments that he is "brutal, vicious, heartless, mean."[22] Similarly purely evil agents, from the perspective of those who affirm the motivational attractiveness of good as opposed to evil, can be criticized in this manner. However, these accusations hardly prove useful. Most agents seem to be disposed towards what is considered the good—that is, whatever they normatively construe to be good. But again, just as one can posit supererogatory agents or "moral saints" on one end of the extreme, there seems to be a kind of agent that can embrace, as the motivating incentive of their actions, a will to do what is evil *because* it is evil.[23] Is the purely evil agent just the result of some conceptual game or is there actual substance to it? Returning once again to the beginning of this inquiry, the question of approach is critical in order to address the issues. It is not just a matter of what an evil agent does that makes him evil, but how and why he arrives at those actions as well.

Though thrill serial killers seem to reflect some features of pure evil, they are not adequate test cases. Instead, some philosophers have turned to literature to cite possible examples. McGinn and Haybron, in their musings on pure evil, both focus on the character of John Claggart in Melville's *Billy Budd*.[24] Unlike other evil characters, his nature is described as being of "natural depravity" and an evil nature "not engendered by vicious training or corrupting books or licentious living."[25] In the novel, Claggart—a high ranking officer—conspires to harm lowly sailor Billy Budd for seemingly no benefit other than just to do it. The calm and calculated demeanor of the man further underscores the quiet yet sinister force that lurks within him. The purely evil agent as such understands the moral gravity of his actions and exhibits a natural preference to do evil. Melville suggests this natural preference, like a scorpion's nature to sting,

---

[22] Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, ed. Ferdinand Schoeman (New York: Cambridge Press, 1988), 271.
[23] Section 3.2 will further explore this notion, sometimes called "the mirror thesis".
[24] McGinn, *Ethics, Evil, and Fiction*, 63 ; Haybron, "Evil Characters," 131.
[25] Herman Melville, *Billy Budd*, ed. F. Barron Freeman (London: Oxford University Press, 1948), 187 [Ch. XI].

attracts Claggart to the malevolence of evil which would repel ordinary men.[26] McGinn rightly recognizes that one cannot simply explain this kind of character as one could the average rogue.[27] Pure evil is often cloaked, as it is with Claggart, behind a convincing façade and operates on an entirely new level than perverse evil.[28]

To stubbornly characterize such evil as simply brute viciousness is to betray the most terrifying feature of purely evil agents: their self-aware, astute intellect. Pure evil agency here has a subtle, almost Machiavellian, nature that undercuts the conventional descriptions of evil as a bestial, impulsive force. Similar to Erich Fromm's earlier warning, individuals should be wary of putting too much stock in the physical identity of evil agency—how an agent looks or appears. This relates to Machiavelli's infamous tract *The Prince* where rulers are advised to use appearances to their advantage. A ruler that can seem to display good and admirable qualities and not have to continually observe them is in the best situation to rule.[29] The purely evil agent may also exercise this ability to control how he/she looks (to others). If an evil agent does not need to be "dressed" a certain way to be evil, then perhaps it is unnecessary for a purely evil agent to manifest all their evil qualities at once or consistently but merely know how and when to make use of them.

---

[26] Melville, *Billy Budd*, 192 [Ch. XIII].
Here is the following excerpt for the reader's benefit: "With no power to annul the elemental evil in him…apprehending the good, but powerless to be it; a nature like Claggart's surcharged with energy as such natures almost invariably are, what recourse is left to it but to recoil upon itself and like the scorpion…act out to the end the part allotted it."
[27] McGinn, *Ethics, Evil, and Fiction*, 64.
[28] It should be noted though that some literary analyses of Claggart's character suggest a rationale for his actions that would undercut this assessment. One could argue, based on the thoughts of Melville's narrator in *Billy Budd*, that Claggart despised Billy's innocence and ethical beauty. That is, Billy was a tortured reminder for Claggart of what he could never be and he sought to destroy Billy to satisfy his own ego. His motives then under this interpretation would be instrumental in much the same way as Hitler's motives to commit genocide were for the sake of an agenda and not a principled stance towards malevolency as such. If this is the case, then Claggart would cease to be representative of pure evil agency as it is defined here. But there is some ambiguity in this interpretation that allows for alternative readings.
[29] Nicolo Machievelli, *The Prince*, trans. George Bull (New York: Penguin Group, 2005), 73-76 [sec. xviii].

The difficulty in clearly portraying real-life instances of purely evil agency should not discount it being a plausible conceptual feature in moral discourse. While the rarity and enigmatic nature of pure evil may yield doubts about its conceptual coherency, literary and/or cinematic sources to the contrary should not be ignored. Perhaps the reason why one cannot adequately identify a purely evil agent in reality is because, unlike the exploratory novel or film, people can only know so much about other's inner feelings and motivations. One's own inner states even seem inscrutable at times. As such, particular serial killers may resemble purely evil agents in some way because criminal psychologists in confined rooms can, given enough time, ascertain their psychological makeup. However, these examples take away some of the exceptional aspects which literary and cinematic cases offer with far more clarity—as much as one can hope to obtain at any rate.

Nonetheless, the multifarious parameters explored here lend a flexible perspective to explore evil's underpinnings. But these remarks about evil are merely a preliminary to a much more thorough analysis of evil—pure evil in particular. One aspect that still needs to be synthesized into the dual notions of evil outlined here is the possible implications on moral psychology. Just as there is a wide variety of character types that can be classified as evil, the psychology of evil agency may not be so simple. Before embarking on the metaethical details of this topic, there are some contentious issues to explore concerning the mind-set or mental state(s) of evil agency.

1.2 "Evil, be thou my Good"

Just as evil can have many faces, one must be attentive to its copious roots. As the previous section indicated, evil can manifest in peculiar ways. What this says about the psychology of evil-doers may equally vary. Michael Stocker suggests that philosophers have far too simplified the cognitive structures within moral psychology, especially when it comes to agents desiring what is bad.[30] It is important to be wary of hasty generalizations. Evil characters cannot simply be pigeonholed into broad attitudinal categories. Thus, to complete these introductory remarks on evil, one must look into its psychological dynamics and what can be construed from the previous section's division of evil into perverse and pure forms.

Many agents, of course, can be explained quite easily. The psychological mechanics of mental neuroses, in terms of childhood development and education, fit very well with numerous perverse serial killers and sociopaths. Others have twisted moral codes, like Hitler, with intermingling cases exhibiting strong archetypal characteristics that stretch the boundaries of perverse evil. In this vein, it was suggested earlier that human emotions can sometimes take on absurd levels of obsession, fear, and anger and push individuals to the extremes of action. Anakin Skywalker's turn to evil could be explained in this way by his single-minded pursuit of personal glory and obsession with losing all that he loves. The source that signified this turning point—Anakin's love of his wife and their welfare—was, for Zizek, something one would normally classify as praiseworthy or good if it had not led Anakin to results that utterly maligned its moral value.[31] Similar characters, such as Captain Ahab and Achilles, could also be explained

---

[30] Michael Stocker, "Desiring the Bad: An Essay in Moral Psychology," *The Journal of Philosophy* 76, no. 12 (1979): 739.

[31] Zizek, *The Parallax View*, 101.
Zizek's analysis of Anakin's development in the *Star Wars* saga is critical of the way producer George Lucas actually portrayed the character. Nonetheless, Zizek's believes that Anakin's transformation into Darth Vader symbolized a turn to evil "through the very wrong mode of his attachment to the Good…becom[ing] a monster out of his very excessive attachment to seeing evil everywhere and fighting

in terms of obsession and fear.[32] Even then, however, there are exceptions which elude straightforward assessment. How does purely evil agency fit into this conceptual scheme?

With evil agency of any kind, many philosophers think that "bitterness and other states of mind constitute a sufficient motivation for immoral action".[33] While this is a strong candidate for explaining all sorts of perverse agents, there are some that think the pure evil model also falls into this psychological rationalization. Roy W. Perrett (among others) suggests that for an agent to be evil, "action has to flow from a particular kind of depraved character".[34] Depravity, however, is a sign of perversion and purely evil agents as described in the previous section are above being driven by instrumental desires as the basis for their motivation to act. That is, pure evil is as much a principled stand as its contrary, the moral saint, who does good for its own sake beyond what is morally essential. Some philosophers go even further and question the coherence of evil-qua-evil motivation entirely. Immanuel Kant in particular argued that it embodied an absurd-like demonic malevolence which can only be possessed by an inhuman creature; that is, such agency cannot be fully instantiated in human beings.[35]

Are evil agents, perversely and purely evil alike, determined by a largely negative moral psychology? And furthermore does this overlapping psychology threaten section 1.1's conceptual

---

it." This brings extra meaning to one of Friedrich Nietzsche's oft-cited quotes from *Beyond Good and Evil* [sec. 146]: "Whoever fights monsters should see to it that in the process he does not become a monster."

[32] Perhaps the white whale, also known as Moby Dick, took more than just Ahab's leg but also his pride. Ahab's single-minded pursuit of the whale can be interpreted as a fervent obsession to regain his wounded pride and peace of mind. It follows that Ahab's willingness to sacrifice all mortal interests—including the welfare of his crew and ultimately himself— represents a perverse obsession fueled by character defect. Similarly, Achilles pursuit of immortal glory in *Troy* (2004) suggests a fear of becoming insignificant—of being forgotten. While the circumstances are different, the characters both embody an obsession taken to perverse levels. Anakin, like Ahab and Achilles, is a character that demonstrates one of the many paths towards [perversely] evil agency.

[33] Samet-Porat, "Satanic Motivations," 85.

[34] Roy W. Perrett, "Evil and Human Nature," *The Monist* 85, no. 2 (2002): 305. Also see McGinn, *Ethics, Evil, and Fiction* (1997); Haybron, "Evil Characters" (1999); and Samet-Porat, "Satanic Motivations" (2007) for similar commentary of evil's nature stemming from predominantly negative features of human emotion.

[35] Immanuel Kant refers to this as the devilish or diabolical will in *Religion within the Bounds of Bare Reason*. This will be explored later in section 2.3 and chapter 3 as a whole.

distinction between perverse and pure evil? In order to preserve this conceptual distinction from collapsing into one all-inclusive category, one must question the hypothesis that evil is principally rooted in negative states of mind. That is, one must explore the possibility that evil can result from motivational elements other than bitterness and hatred. These things are largely deemed instrumental features of moral motivation and not befitting of purely evil agency. Additionally, the psychological dynamics of intentional evil must be clarified with a distinctive example of evil-qua-evil motivation.

The debate over Milton's Satan as a paradigm case of evil has been a noticeable object of discussion over the years.[36] And it is a suitable example to test the claim that evil agency is rooted in negative and corrupting psychological states. The iconic phrase "Evil, be thou my Good"[37] has inspired discussion about whether Satan is really an evil agent or actually a closet lover of what is good. This is important in that it is an attempt at genuinely depicting the intentional attitude of evil agency. If Satan and other evil agents are identifiable as closet lovers of what is good, then their characteristics and subsequent mental states are nothing but (negative) reactions towards the existent moral code. Evil becomes synonymous with mere deviancy, something indicative of a rebellious teenager. In other words, agents who are believed to exemplify or embrace evil are rather perverse agents redefining what (they think) is good to their own purposes—perhaps subconsciously.

Considering the storyline of *Paradise Lost*, there are many philosophers that are sympathetic to this interpretation. "Evil, be thou my Good" is uttered by Satan after rebelling against God, losing, and as a result being banished forever to his own hell. His emotions—bitterness, rage, pride—are pouring out as he comes to grips with the situation. It is at this point

[36] See Elizabeth Anscombe, *Intentions 2nd ed* (London: Harvard Press, 1976); David McNaughton, *Moral Vision* (Oxford: Blackwell Publishing, 1988); Robert Dunn, "Is Satan a Lover of the Good?", *Ratio (new series)* XIII, no. 1 (2000); and Samet-Porat, "Satanic Motivations" (2007).
[37] John Milton, *Paradise Lost* (Oxford: Oxford University Press, 2005), 108 [Bk. IV: 110].

that Satan gathers his thoughts and bolsters his resolve. His future aims are attuned to the pursuit of evil as an adversary to God.

Anscombe suggests that Satan sees his evil as a good in the sense that good is the "intact liberty in the unsubmissiveness of [his] will."[38] In other words, Satan by revising his moral aims has created his own sense of good in defiance of God's good. He seems to acknowledge from the standpoint of morality, or specifically in *Paradise Lost*, from God's moral perspective, that his defiance is contrary to what is deemed good among the rest of God's creation. But McNaughton points out that Satan—if he is a closet lover of the good—would be using the term "evil" descriptively rather than in a normal evaluative way.[39] That is, he is using "evil" in an inverted-commas sense. Essentially the declaration "Evil, be thou my Good" can be interpreted as semantic wordplay. McNaughton, along with Anscombe, would say that Satan does not pursue evil *because* it is evil.[40] Just as Anakin Skywalker declares to Obi Wan in *Revenge of the Sith* (2005) that "From my point of view the Jedi are evil,"[41] the suggestion above is that Satan views himself, his actions, and motivations as good even though he uses the label of "evil". He exalts evil not for what it *is*, but for what it represents for him. Defiance is the means by which Satan confirms his immovable pride and unsubmissiveness. This is what Anscombe and McNaughton mean when Satan is construed as a closet lover of the good.

While this is something one would expect of a perversely evil agent, the purely evil agent is someone who is evil precisely because of the evaluative meaning of evil itself. Milton's Satan does seem to embody the bitterness and negativity linked with perverse qualities of evil. As mentioned previously, some are sympathetic to this interpretation of Milton's Satan, largely because of the specifics of the narrative. The details surrounding Satan's motivations seem

[38] Anscombe, *Intentions*, 75 [sec 39].
[39] McNaughton, *Moral Vision*, 136.
[40] Ibid., 141-142.
[41] *Star Wars: Episode III - Revenge of the Sith*, directed by George Lucas (Marin County, CA: Lucas Films, 2005), DVD.

overly tied down to the circumstances of his fall. The sting of Satan's recent defeat and his obdurate pride are contributing factors to the persona presented by Milton and lend support to Anscombe's account at face value. Similar reasoning led Samet-Porat to be wary of appealing to Satan as an exemplar of evil.[42] Anscombe and McNaughton's interpretation, however, is not indisputable. Strong reasoning for a different interpretation of Satan's intentions is available—one that does not support him as a closet lover of the good.

When Milton's Satan states "Evil, be thou my Good", what could he have been promoting or pursuing? What were his intentions? As Anscombe already suggested, one could interpret Satan's evil as redefining his own view of good in direct opposition to God—similar to how an ordinary villain could view his actions in relation to the "goodie-two-shoes" hero.[43] That is, he is a closet lover of the good. He pursued evil because it embodied the means for expressing insubordination to God. What was evil from God's perspective became his new good. The underlying premise here is that Satan must see these intentions as good in some way and thus "evil" is a mere placeholder, an empty term no longer having any original content. Good for Anscombe is multiform and all that is needed to confirm Satan as operating under the aspect of good is his intention of "wanting" it.[44] But this view of evil intention makes it impossible for an agent such as Satan (or any agent at all for that matter) to pursue evil and simultaneously desire evil for itself. Thus, there seems to be no such thing as a *positive* expression of evil intention; it must manifest indirectly as a negative reaction (e.g. bitterness, hatred, etc). This conclusion

[42] Samet-Porat, "Satanic Motivations," 85.
[43] For example, consider how many comic-book villains often come from relatively well-intentioned backgrounds and simply do not start out as corrupted wrongdoers. The problem is that when their good intentions go awry or are stifled, they can turn towards extremism. "Doc Ock" in *Spiderman 2* (2004) had the well-intentioned goal of scientific progress, creating a new energy source, but his initial failure only proved to drive him to accomplish the goal by any means necessary. People viewed his actions as evil, for threatening the lives of others. But we easily forget what Doc Ock, perhaps like Milton's Satan if Anscombe's interpretation is correct, might have thought of himself and those fighting against him. One person's "evil" can easily be another's "good". Yet it does not necessarily follow that Satan must see his actions as *summum bonum* (Latin for "highest good") and not some other instantiation of "good".
[44] Anscombe, *Intentions*, 75 [sec 39].

though seems premature, as if there is not a corresponding interpretation available for reading Satan's anthem "Evil, be thou my Good" in a way that supports purely evil motivation.

There are some questionable portions of Anscombe's account that warrant attention. First, as previously mentioned, Satan's pursuit of evil is done within the moral framework of what God deems good and bad (or evil). But then what sense can be made out of the idea that Satan is redefining the good for himself? Like most perversely evil agents, it is argued that Satan twisted good and evil to serve his own purposes. This is not the same thing though as redefining a new conception of the good if it is parasitic on an existent moral framework. Robert Dunn likewise points out that "Milton's Satan does *not* have a substantive theory of the good under which…the actions that promote them count as goods."[45] In other words, Satan does not have a stand-alone theory of moral goodness and badness. This opens up the possibility that Satan could have merely reversed the direction of his moral compass to evil *for itself* instead of being a closet lover of the good. The mere fact that someone wants some particular end does not in itself demonstrate that the agent thinks it is good. On the contrary, there is evidence that may suggest Satan does see his aim(s) as bad or evil in the normal evaluative sense and the term "good" within the phrase "Evil, be thou my Good" is being utilized in a different way.

Before uttering that iconic phrase, Satan declared, "So farewell hope, and with hope farewell fear, Farewell remorse: all good is lost to me."[46] This short passage preceding Satan's anthem can be interpreted differently. If Satan aligns himself to evil with the thought(s) that there can be no hope in what he does, then by what means can he be construed as a closet lover of the good? Anscombe's working assumption is that one presumably wants/prefers the good because it confers some perceived benefit or value. And in this vein, to prefer something over another is to

[45] Dunn, "Is Satan a lover of the good?", 15.
[46] Milton, *Paradise Lost*, 108 [Bk. IV: 100-110].

make a "good" of it. But Stocker makes a strong case against such a blatant assumption.[47] Not only could the believed good (as opposed to a perceived good) fail to attract Satan—as seems to be the case with the previous quotation—but he could be attracted instead to the actual, believed bad. Stocker maintains that "only against a certain assumed background of agent mood and interest does citing the believed good make an act intelligible."[48] The background of Satan's moral disagreements does not seem to stem at all from a difference in perceived goods (i.e. "My good" vs. "God's good"). But rather, Satan's moral disagreements are rooted in the motivational inefficacy of the believed good; he is motivated to do evil precisely because of the fact that it is not good. In this sense, Satan's pursuit of evil takes the place of what would ordinarily be a person's pursuit of good but there are no illusions—as with perversely evil agents—that one is committed to evil as evil. A closer look at *Paradise Lost* may possibly unhinge Anscombe's and McNaughton's position in this way.

If it is the case that Milton's Satan does not substantively revise moral goodness and badness but only reverses his aim(s) from good to evil, then one must re-evaluate the psychology of evil intentions in light of the details to determine whether Satan could conceptually be a purely evil agent—positively affirming evil *for itself*. Before Eve is tempted in the Garden of Eden, Milton presents a self-reflective moment from Satan as he infiltrates Paradise: "But neither here seek I, no nor in heaven To dwell, unless by mastering heaven's supreme; Nor hope to be myself less miserable By what I seek."[49] The mood and interest of Satan here seems contrary to any label of closet lover of the good. He not only recognizes that his intention to do evil will do nothing to cure his misery, but also accepts and anticipates the outcome with the utmost of his being. He sees himself as his own master rebelliously renouncing moral goodness, but not to the extent that he reassigns his moral values. Rather than a reassignment of what is good and evil,

[47] Stocker, "Desiring the Bad: A Lesson in Moral Psychology," 738-753.
[48] Ibid., 746.
[49] Milton, *Paradise Lost*, 251 [Bk. IX: 125-130].

Satan appears to align his motivational priorities inversely against his moral judgments without changing or twisting the moral source. Otherwise, Satan's expressions of pride and envy would not make sense if he were a perversely evil agent with a twisted sense of self-righteousness on the same level as Hitler. The triumph over God's newly created project, Adam and Eve, is Satan's demonstration of evil's pure power against the good. He has went "all-in" with the motivational shift from good to evil and, in Thrasymachean fashion, wants evil to outdo good in this new war against God. As much as Satan may seemingly derive his evil from "negative" sources, he expresses his pride, bitterness, and envy positively for itself rather than reactively.

Given these new developments, Dunn rightly suggests "Evil, be thou my Good" could be understood as Satan making evil his new "criterion of success for any related action" instead of "targeting" [and redefining] a value or good.[50] Milton's Satan could consider the evil in defying God as good in the sense that one can generally consider the success of performing a designated task as good. For example, if a person aims to carry out a bank heist, he/she can judge the success of that action as good without having to normatively reassign his/her understanding of moral goodness. In other words, a hypothetical bank robber can very well know and operate within an existent moral code that condemns bank robbing, but have aims which motivationally dispose him/her to do what is bad.

Thus, when Milton's Satan enshrines evil as his new good, there are other ways to interpret the use of "good" besides normatively. The statement, "That was a good bank robbery I did," can equate to, "I did a good job robbing that bank," rather than a commitment to the normative value judgment, "I think robbing a bank is a (morally) good thing to do." Dunn would label the former statement as invoking, "a formal sense of 'good', in which any aim whatsoever

[50] Dunn, "Is Satan a lover of the Good?", 15.

counts as good" as a matter of success.[51] This sense of good is far removed from the good being evoked by Anscombe and McNaughton—an evaluative good.

McNaughton, like Anscombe, thinks Satan's reflections on the good lost to him reveals his motives to do evil. Satan chooses the path of evil because "it represents [his] only remaining hope of power."[52] Exiled, cursed, and with a wounded pride, Satan resorts to evil as a tool to exalt his unsubmissiveness and affirm his egocentric defiance. This interpretation, though, places undue weight on good to the extent that to *will* anything is to desire the good in some way. One should, naturally, be able to account for genuine cases where the agent's circumstances and motivational background suggest a twisted conception of the good (e.g. Hitler); but it should not be set as the default condition for every agent who utilizes the term "good" in a moral situation. Not only does "good" lose its substantive meaning by being tautologically subsumed under "whatever one positively wills," but also ordinary moral discourse is turned on its head.

There is a conceptual gap in Anscombe and McNaughton's interpretations that does not adequately match the psychological force of the term "evil" in moral discourse. When said with tremendous vigor that "X is positively (or purely) evil," it is not always meant to denote that X is operating under a twisted conception of the good. On the contrary, X could be operating under an entirely different framework than any common moral agent. As such, the purely evil agent's motivational disposition operates in the reverse embracing evil because of its "evilness" instead of evil because of some (falsely) perceived good in the act. While certain conditions are required to be purely evil and these conditions are both intricate and rare, there is no reason at this time to reject the conceptual plausibility of pure evil. Not all evil agents possess the same mindset and it is the mindset and intentions of an agent (including emotions) that ultimately distinguishes perverse and pure evil.

---

[51] Ibid.
[52] McNaughton, *Moral Vision*, 142.

Satan is surely under no illusions. He knows his aims are evil and that they are detrimental to his condition (i.e. will not make him any less miserable). As such if there is any good that he seeks, it is the success in action of his evil aims rather than reinventing good under the label of "evil". Most people may feel some kind of cognitive dissonance or mental pangs of guilt for actively doing what they know and believe to be wrong. But certain individuals, like the character of Satan, are exceptions that show it is possible to positively acknowledge evil for what it is and perform it regardless. Samet-Porat has similar reservations about Anscombe and McNaughton's account(s) of Milton's Satan, but is hesitant about treating Milton's Satan "as the paradigmatic satanic agent."[53] In fact, she seems more inclined toward the example of Richard the III (via Shakespeare's depiction) and the peculiar case surrounding Nazi propagandist Joseph Goebbels. However, Samet-Porat does also insist that not all evil should be rendered perversely evil, to use "evil" in an inverted-commas sense.[54]

Granting the philosophical and theological baggage of this character, there can at least conceptually be positive expressions of intentional evil. One should not presume evil fits any particular label or identity, even the most commonsensical; for one now has reason to think that some evil agents are not possibly determined by a negative moral psychology. Bitterness and negativity on the whole are nevertheless pervasive contingent features of evil. These conceptual and psychological insights advance our inquiry into the metaethics of evil by demonstrating the sheer elasticity of evil's framework. Such open-endedness, though, will prove to further mire this inquiry with challenging questions once the intricacies of agent motivation and judgment are under consideration.

[53] Samet-Porat, "Satanic Motivations," 85.
[54] Ibid., 93.

1.3 Developing a Metaethical Understanding of Evil

Section 1.1 identified and examined perverse and pure conceptualizations of evil. This greatly expanded the parameters under which agents can be assessed as "evil". Additionally after considering the complex underpinnings of evil's moral psychology (with characters such as Milton's Satan) in section 1.2, these dynamics need to be taken into account within an overall metaethical understanding of evil. That is, surveying the moral and psychological landscape of evil establishes groundwork for analyzing metaethics appropriate to this thesis. The previous section proposed some interesting conclusions about purely evil agents—that such characters with evil-qua-evil motivations are conceptually plausible and they perhaps need not be rooted in negative psychological states. But, this conceptual depiction of evil is incomplete without some corresponding metaethical context to make sense of it. How is one to understand the inner workings of an evil agent's motivations in relation to their judgments and actions?

The metaethics relevant to this thesis involves two opposing philosophical positions: internalism and externalism. The focus will be only on a specific type of internalism and externalism: motivation internalism/externalism. Nonetheless, these frameworks have further uses for a number of philosophical disciplines. In epistemology, for instance, internalism about justification asserts that justification for beliefs is necessarily derivable from internal factors like the mental contents of an agent's consciousness. In philosophy of language, the debate between internalism and externalism arises from questions about the origin of semantic meaning—whether the meaning of a word is cognitively (i.e. internally) construed or rather determined externally by environmental conditions.

There are a few variations of internalism/externalism in metaethics. Bernard Williams is well known for highlighting internal and external reasons for action—otherwise known as *reasons* internalism and *reasons* externalism—which explores the relation between one's moral

judgment and the justificatory reasons for that judgment.[55] The internalism and externalism under consideration here is closely related to—but regardless logically independent of—Williams' distinction.

Motivation internalism/externalism explores how an agent's evaluative judgment that "It is morally good or not morally good to do X" relates to any subsequent motivation to act in accordance with that judgment. Suppose an agent expresses the judgment "It is not morally good to take another person's possessions" or more specifically "I should not steal that person's wallet" during a moment of moral deliberation. By expressing this judgment about his/her moral belief(s) about some matter, the agent presumably considers the judgment authoritative and action-guiding. Whether the agent's moral beliefs are considered propositional claims—statements capable of being either true or false—or mere emotive expressions[56] is an important issue to consider but not relevant to the aims of this inquiry. Regardless of whether cognitivism or non-cognitivism is the correct metaethical position with regards to moral beliefs/judgments, they can (perhaps) occupy both sides of the debate between motivation internalism and externalism.[57] So, for the sake of brevity, this inquiry will focus exclusively on the competing theses of motivation internalism/externalism and how evil can be used as counterexamples to either support or refute each motivational framework.

[55] See Bernard Williams, "Internal and External Reasons," in *Moral Luck*, (New York: Cambridge Press, 1986).

[56] One rather explicit example of non-cognitivism in metaethics can be found in A.J. Ayers, *Language, Truth, and Logic*, (New York: Dover Publications, 1952), 102-120. For Ayer, (evaluative) moral judgments are merely expressions of emotional approval or disapproval. The judgment "Stealing is wrong" is the equivalent of saying "Boo stealing!" In other words, ethical judgments as expressions of value are not capable of being truth or false. This position, commonly known as emotivism, is one popular instantiation of non-cognitivism.

[57] See Richard Joyce, "Expressivism and Motivational Internalism", *Analysis* 62, no. 4 (2002). He makes a strong case that expressivism need not imply, or commit oneself to, motivational internalism.

The motivation internalist posits a necessary[58] or non-contingent connection such that when an agent affirms the judgment about stealing being wrong it is—ceteris paribus—sufficient enough to be motivating to act. In other words, under motivation internalism, the sincere approval of one's moral judgments establishes a motivating element within the judgment itself. To borrow Richard Joyce's phrasing in *The Myth of Morality*, motivation internalism is the view that "it is *necessary* and *a priori* that any agent who judges that one of his available actions is morally obligatory will have some (defeasible) motivation to perform that action (emphasis in original)."[59] The underlying premise here seems to be that moral judgments by definition contain an implicit "must-be-doneness". As such, an agent's moral judgments serve to broadcast one's motivational dispositions. When someone forms a moral judgment but then at the next moment indicates motivations to do the contrary, the internalist view quite easily pinpoints the source of the bewilderment; there is some connection here that agents generally have between their judgments and subsequent motivations to act on them.

Whether the agent successfully acts according to his/her judgment is irrelevant to the existence of said motivation. For the internalist, the agent's judgment merely lends sufficient motivation for the agent to *want* to act on it. If a moral judgment fails to even minimally affect the decision of the agent, then one could plausibly question the validity (in terms of sincerity) of the judgment; as a result internalism would remain unblemished. A general outline is provided (see Figure 1 below) to demonstrate the motivation internalist framework in two stages; the first stage is within the domain of an agent's deliberation and the other stage an agent's motivation to act as a result of that assessment.

[58] There may be some metaphysical baggage with this term and major differences in how philosophers interpret it. For future reference, these terms are utilized strictly in the logical sense—contingent vs. non-contingent.

[59] Richard Joyce, *The Myth of Morality* (New York: Cambridge University Press, 2001), 18.

Stage One (S1)	S1: An agent (A) forms a moral judgment (J) if A judges and sincerely affirms J such that J has the property of being an action-guiding moral belief.
Stage Two (S2)	S2: An agent (A) is motivated to act on a judgment (J) if A forms J such that J establishes motivating reasons to act in some situation (S) which A sincerely affirms as morally required.
Motivation Internalist Thesis (MIT)	MIT: Given the transition from S1 to S2, there exists a conceptual relationship betweeen approving a moral judgment and being motivated in some way to act on that judgment.

Figure 1

The above staging together represents the motivation internalist thesis (MIT). At the basic level, it is a working model that proposes to make sense of the motivational profile of human beings when making moral claims. What does it mean when an agent expresses a moral judgment about some state of affairs? Simply, for the internalist, an agent's definitive moral judgment seems to also establish the agent's motivational disposition to act in that way—even if that motivation is not forthcoming and ultimately fails to spur the agent to act in that manner.

In other words, if an agent sincerely forms an evaluative moral judgment and subsequently encounters a situation where the judgment demands a course of action, then the agent has sufficient motivation to be inclined towards the execution of that judgment in one's actions. If someone is not motivated by their moral judgments and as a result fails to be moved by them in the context of a situation, then the internalist can critique the conditions surrounding

the person's judgments or beliefs and explain the disconnect in terms of ignorance, insincerity of judgment, irrationality, or perhaps even mental defect.[60]

On the other hand, the motivational externalist, as Shafer-Landau puts it, "endorse[s] the conceptual possibility of one who sincerely makes moral judgments but is entirely unmoved by them."[61] In other words, the externalist does not think that there is a necessary and a priori connection between approving of a moral judgment and having the motivation to act on it. Rather, the connection is contingent on external factors which may override how agents are motivated by and/or relate to their pronounced moral judgments. As such, considering the very possibility that an agent's judgment and motivation to act can be disassociated makes one partial to an externalist view. Contrary to internalists, externalists do not regard judgments as *necessarily* motivating even in the sense of an agent wanting to act on it. Similarly, Robert Lockie describes motivation externalism as merely doubting the essential motivating influence of judgment as opposed to the extreme view that judgments are never (internally) motivating.[62] Just as internalists grant that motivations can fail to develop into the appropriate action, the externalist need not deny that judgments often have motivating force for agents—only that this does not, thereby, establish a necessary conceptual relation between the two. As such, the internalist framework outlined above in Figure 1 only offers one of presumably many thought processes by which an agent can be motivated to act.

Even though the externalist criticizes the transition from S1 to S2 in Figure 1, this does not prevent the externalist from potentially agreeing with some of the internalist explanations for particular cases. There are likely legitimate instances where an agent is being insincere with his/her judgments or perhaps has schizophrenia or a psychological defect of some kind that severs the strong connection between agent judgment and motivation. But the real question to begin

[60] Some of these explanations will be explored with more detail in Section 2.1.
[61] Russ Shafer-Landau, "A Defense of Motivation Externalism," *Philosophical Studies* 97, no. 3 (2000): 271.
[62] Robert Lockie, "What's Wrong with Moral Internalism," *Ratio (new series)* XI, no. 1(1998): 25-26.

with is whether the connection is strong enough to be considered necessary and a priori. Many partial to the externalist view attempt to devise agent counterexamples that try to refute the MIT.[63] Amorality is one such counterexample where an agent simply withdraws from morality altogether while still sincerely professing knowledge of moral rightness and wrongness. The complexities of this debate extend among numerous philosophers with a noteworthy paper trail of publications.[64]

The conceptual and psychological conditions of evil that were outlined in the previous sections concern these metaethical positions in more ways than one. This inquiry, though, will restrict itself to the disputes between motivational internalists and externalists. Many ethical systems contain basic axioms that depend on such metaethical foundations. A triumph for either position could significantly change the moral landscape as it is understood by (normative) ethical theory. Platonic and Kantian ethics, for instance, are predominantly internalist whereas most sentimentalist and non-intellectualist moral theories tend towards an externalist view (e.g. Hume). There are discrepancies, though, among internalist structures that may significantly shape how evil is construed. Pure evil agency at face value seems like a strong counterexample to the internalist thesis because it suggests the existence of an agent that can judge or deem something to be morally good but in turn be motivated to do its contrary, evil. In order to better understand

[63] To list a few recent examples (some of which will be revisited later): Robert Lockie, "What's Wrong with Moral Internalism" (1998); Christian Basil Miller, "Motivational Internalism," *Philosophical Studies* 139, (2008); Jon Tresan, "Metaethical Internalism: Another Neglected Distinction," *Journal of Ethics* 13, (2008); Antti Kauppinen, "Moral Internalism and the Brain," *Social Theory and Practice* 34, no. 1 (2008); and Andrew Sneddon, "Alternative motivation: a new challenge to moral judgment internalism," *Philosophical Explorations* 12, no. 1 (2009).

[64] To describe a few in detail: McNaughton, *Moral Visions* (1988); Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994); and Joyce, *Myth of Morality* (2001). McNaughton presents a broad and accessible survey of the metaethical field, though lacking recent externalist literature; he seems to support the internalist position. Smith offers a unique defense of the MIT in the form of modeling an ideal rational agent, which in itself has elicited numerous responses and criticisms. Joyce, in the first few chapters of his book *The Myth of Morality*, adds a moral error theorist twist to the internalist/externalist debate arguing that motivation internalism is at the heart of the error within moral discourse.

the inner workings of evil, this apparent discrepancy needs to be further clarified by motivation externalist counterexamples and coherently examined by motivation internalist accounts.

Rather than get buried in a mountain of unnecessary details, important though they may be, the primary aim here is to specifically concentrate on internalist accounts of evil and their respective externalist objections. In other words, the point is not to undertake the daunting task of concluding whether moral judgments are necessarily motivating (or not), but rather to investigate internalist views of evil in light of the conceptualizations highlighted in sections 1.1 and 1.2. On this metaethical level, pondering the relation between agent judgment and motivation could advance philosophers' normative understanding of evil.

The following are just a few of the questions this inquiry will examine in the subsequent chapters: How does the motivational internalist explain purely evil agents, such as Milton's Satan, who seem to acknowledge the good but are motivated to do the contrary? How does motivation externalism compare with internalism? That is, how do external factors influence agent motivation and/or judgment for both perverse and pure evil agency? And lastly, what are the components driving these metaethical positions that support the idea of evil?

1.4 Benchmarks

Before examining internalist and externalist reasoning, it would be helpful to establish suitable criteria or benchmarks that keep both theoretical positions at a fair starting point in the discussion. While both theoretical camps work directly against the other, this inquiry will impartially mediate their strengths and weaknesses on a level playing field. An argumentative point sometimes forwarded, which must be dispelled, is that the mere existence of amoral or akratic agents satisfies as proof against the internalist position. In other words, the externalist is said to have a certain advantage over the internalist in terms of the burden of proof.[65] It is tempting here to suggest the same about evil agents, thereby automatically putting internalism on the defensive "hot seat".

While the externalist position can be described as the skeptical alternative to the internalist thesis, the reader should for the time being set aside the question of which position has the burden of proof. Even if, as some externalists insist, internalism makes the positive claim,[66] one should not impose such strict conditions at the outset. The motivation internalist should not need to *comprehensively* explain every possible counterexample in order for their explanatory model to be considered. Similarly, externalists should not be expected to prove a negative and disprove every single aspect of an internalist account. There needs to be realistic, yet challenging, objectives to fulfilling the task of explaining evil. Additionally, there must be something that, if achieved, would be a reliable indication of a position's success or advantage.

Any benchmark must begin by safeguarding this inquiry from falling prey to rationalizations that try to beg the question. To use a popular example, it would be circular to argue from the premise that "The Bible is God's word and everything in it is true" to the conclusion "Therefore, God exists" because the premise is assuming the truth of the conclusion

[65] See in particular Shafer-Landau, "A Defense of Motivational Externalism" (2000).
[66] Ibid., 271.

when it should be trying to demonstrate the conclusion. If an internalist or externalist model of evil were to radically change the way in which our moral discourse reflects on the concept "evil", then there should be good explanatory reasons to explain this discrepancy—absent of ad hoc reasoning. Otherwise, it should be able to reasonably account for the reality of evil in life and in moral discourse. This would very much include atypical agents, like Milton's Satan, who seemingly comprises an entirely distinctive category of evil that was explored in section 1.2.

Whatever mechanics of evil proposed or moral psychology posited should address the reality of evil's *multifarious* manifestations. As section 1.1 outlined with some persuasive force, evil can have many faces and roots. That is, there is more than one way to be evil and the details matter. Both the internalist and externalist need to address or keep in mind the diversity of evil which pervades our moral discourse—from the poetic musings of Milton's *Paradise Lost*, the literary creation of Claggart in Melville's *Billy Budd* to the recent cinematic ambiguities of the *Star War* saga's Anakin Skywalker and The Joker in *Dark Knight* (2008). The following is a tentative outline of the core benchmarks that in later chapters will be explored in some instantiation of internalism:

- B1 '*Relevance to Moral Discourse*': The account should present *substantive*, positive, components that illustrate how evil operates within the context of agent judgment and motivation.

- B2 '*Explanatory Power*': The account should adequately explain the uses of "evil" in moral discourse without alienating ordinary intuitions about its meaning. If it does run counter in some ways, then it should be able to explain these discrepancies and not resort to ad hoc tactics.

- B3 '*Consistency*': The account should address and reasonably give account of the conceptual distinction fit together with the initial chapter's conceptual analysis and

psychological analysis of evil. If the account is mistaken, then the account should suggest an answer as to the diversity and type of evil agency.

These benchmarks will not be directly emphasized in this inquiry, but will provide direction for the issues at hand. Exploring the motivation internalist and externalist debate will lend a better understanding of evil and its role in moral discourse.

CHAPTER II

Exploring Motivation Internalism and Externalism

This chapter will be surveying the basic argumentative structures that drive the debate between motivation internalism and externalism. At the same time, it will be suggested how each metaethical position conceptually accounts for evil. There are many layers to the debate, however, that must be clarified. The predominant focus among internalists and externalists in the philosophical literature centers on the amoral agent. How the internalist and externalist commonly explain instances of amorality as well as akrasia (i.e. weakness of will) will be valuable to know for later reference. When the evil agent is put into perspective within the context of internalism and externalism, there may be a general pattern of reasoning that one can identify alongside the amoral and akratic agent. While amorality, akrasia, and purely evil agency are completely different conceptualizations, externalists will utilize each in similar ways as critical counterexamples to the MIT. Similarly, internalists have a stockpile of ready-made answers for mostly every deviation from the motivational norm.

While part of this chapter will be devoted to making sense of internalist and externalist reasonings on these related matters, this inquiry is ultimately focused on how evil is understood in relation to agent judgment and motivation—specifically purely evil agency. Andrew Sneddon, among other philosophers, will be cited in the attempt to demonstrate the sort of problems evil—specifically purely evil agency—triggers in the debate. Immanuel Kant's internalist account of evil in *Religion within the Bounds of Bare Reason* will stand out as the most intriguing, if not the best suited, for providing an answer to the inner workings of the evil will.

Section 2.1 will explore amoral and akratic agency in terms of how it has shaped the debate between motivation internalists and externalists. The MIT has been frequently modified as conceptual and empiricial information arises in moral discourse. To illustrate the significant progress made by contemporary internalists, Plato's model will be used to show critical problems with some internalist formulations regarding how rationality is understood with theories of what is good. Plato's assumptions about evil agency are not beyond question and have, to some extent, been discarded by modern-day internalists. In section 2.2, the discussion will turn to how the conceptualizations of evil previously outlined in the first chapter can influence the debate. The purely evil agent can be presented as an additional counterexample for externalists to use against the MIT and some internalist answers will struggle making sense of purely evil agency. After assessing how evil agency fares alongside other (amoral and akratic) counterexamples, section 2.3 will focus on introducing the Kantian model. Immanuel Kant's account of evil evades some critical problems that will be discussed in this chapter but at the cost of rejecting the possibility of pure evil for human beings.

2.1 Amorality and Akrasia

In moral discourse, it is a common notion that people's conduct or actions are to some extent guided by their moral judgments. Moreover, judgments are generally construed as expressions of belief—the things that individuals and groups identify with. The debate between motivation internalism and externalism explores whether moral judgments themselves have motivational efficacy or the source of motivation is largely external to moral judgments. That is, the issue involves the degree and the extent by which motivation enters into judgment once it is consciously and intentionally uttered by a person (or agent). While there are differences between internalists and externalists within their own theoretical positions, there are core elements that distinguish both.

Motivation internalists generally posit that there is a necessary and a priori connection between believing or judging X is good and the motivation to X. Externalists, on the other hand, argue that the connection between our judgments and motivation is—at best—contingent due to certain counterexamples refuting the internalist picture of motivation. On their view, it is possible to sincerely express moral beliefs and not be motivated to act on them. Much of the debate in this regard has typically focused on externalist counterexamples of amoral agency and akrasia. This section will briefly summarize the progression of this debate in preparation for a new set of counterexamples involving evil agency.

When a moral agent has the belief that "Being kind is a morally good thing to do" and makes a judgment based on this belief in particular circumstances, one generally expects that this judgment will cause the agent to act on it or to some extent be action-guiding. Beliefs are supposed to inform agents how to live by demarcating (what the person deems) right from wrong or good from evil. In what sense then can one have a belief or judgment about something but not be motivated or compelled to act on it? Suppose someone recognized the moral legitimacy of the

previous judgment about kindness. It would be strange if that person then turned around and acted rudely to others with deliberate impudence. One might think the agent misunderstood what he/she was saying or perhaps the agent temporarily lost his/her faculties (via a mental disorder or a traumatic situation that could explain such seemingly erratic behavior). An internalist might say—among other things—that the agent did not sincerely have the belief to form the judgment in the first place. This seems to suggest that belief must be what William James labels a "live hypothesis" and not some lifeless proposition.[67] For an agent to form a judgment, and do so sincerely, there must be some internal appeal; believing in something means having some "willingness to act at all" in that manner.[68]

An externalist, on the other hand, would insist that it is not altogether strange for agents to express beliefs or make judgments and not have any motivation towards implement or acting on them. Recognizing that there are agents suffering from "weakness of will" (i.e. akrasia) is one way to illustrate how motivation is perhaps not completely internal to judgments themselves. It is a case where the agent fails to carry out judgment into demonstrable action because of some overriding influence. Thus the question of whether an agent must judge *and* be motivated by a judgment (e.g. "Being kind is the morally good thing to do") when the occasion calls for doing it, remains open to externalist scrutiny.[69]

Nevertheless, externalists attempt to utilize counterexamples of agents that have deviant motivational compasses but also have the wherewithal to come to the same conclusions as

[67] William James, "Will to Believe," in *Essays on Faith and Morals*, ed. Ralph Barton Perry (Cleveland: Meridian Books, 1962), 33.
[68] Ibid., 34.
[69] As it will be shown later, not all internalist accounts are challenged by akrasia. In fact, some have no problem with integrating the phenomenon of moral weakness and/or indecision into or alongside the domain of immorality. Kant's internalism, for instance, enumerates evil within three grades and one of these—the frail agent—is most likely representative of akrasia. If akratic agents are defined as those who hesitate or have their moral understanding stifled in the midst of action, then, for Kant, there must be some form of self-deception at work. Lawrence Pasternack, in "Can Self-Deception Explain 'Akrasia' in Kant's Theory of Moral Agency?" (1999), expresses some doubts about the possible shortcomings of this appeal. But, at the very least, the attempt to explain akrasia via Kant's theory of agency or some other option has not been rendered impossible.

ordinary moral practitioners about kindness and other moral judgments. Some of these agents may show symptoms of akrasia or be radically amoral in that they are unaffected by the motivational force of their own moral judgments. The recent cinematic portrayal of the Joker in *The Dark Knight* (2008) represents an, albeit extreme, amoral agent with a powerful intelligence and dedication to anarchy. He seems quite aware of society's rules and morality itself, but disregards them willfully. Not all amoral agents fall into this particular lifestyle. Other noticeably amoral agents can have little to no impulse towards criminality—like hermits, recluses, or vagabonds. At the very least, amorality involves a motivational inertness but not at the expense of diminishing a person's knowledge of the matter. Such a person can express a judgment very sincerely but simply does not have the relevant motivation to act on it.

While amoral agents withdraw from moral concerns or questions, they can still possess moral knowledge in the same way one can possess knowledge about proper dinner etiquette. There are two ways in particular that these conditions tend to arise. The agent was raised and instructed in moral discourse from childhood but, similar to how some people lose faith in their religious tradition during adult years, the agent later rejects those norms. Or the agent could have been raised with a great understanding of morality but just never really cared about being a moral person as deemed by social norms. The former can represent nihilists, certain deviant criminals, and perhaps misanthropic hermits; the latter seems to fit the serial killer paradigm addressed earlier in section 1.1.

An example might make this externalist criticism clearer. Suppose a man is attending the figure skating competitions during the next Winter Olympics at the behest of his mother. She is an avid fan of the sport and has immersed him in the practices of professional figure skating as if it was a second language. And like any other regulated activity, there are rules or normative

standards[70] by which one can judge the different kinds of skating maneuvers and the quality of their execution. But while her son never really had the slightest motivational investment in skating, he nonetheless learned the parameters of the sport. He knows what certain terms mean and correctly applies them to various skating activities. As a result, the man is in a similar position to the amoral agent. He has extensive knowledge of a particular discourse and the capacity to form normative judgments according to the standards of this discourse, but—unlike his mother— he is not moved in any motivational sense by his judgments.

At the upcoming Olympic skating competitions, imagine that the man observes one particular skating performance and, based on the established rules of the sport, states that "The skater poorly executed that triple lutz jump." His mother chimes in and forms the normative judgment that "The judges should penalize the skater for switching to the inside edge on that lutz jump." The man affirms this judgment[71] but the mere affirmation of this normative judgment does not seem to be enough to infer the necessity of his motivation to act on it. One could object along the same lines as the internalist's sincerity qualification outlined earlier and suggest that this person is situated as a passive observer and less of a basis to think that their judgments are legitimate. In other words, his judgment fails to be expressed sincerely to count as action-guiding. This would, once again, take the issue back to the question in section 1.2 of whether an

[70] See the section on "The Standards of Evaluation" in J.L. Mackie, *Ethics: Inventing Right and Wrong*, (New York: Penguin Books, 1977), 25-27. In that section, Mackie describes how one can take normative standards as action-guiding but at the same time deny the objectivity of said standards as it relates to moral realism. The standard of, say, grading apples in the food industry is non-arbitrary because there are reasons that drive the standard one way versus another. But Mackie argues that "something may be called good simply in so far as it satisfies or is such as to satisfy a certain desire; but the objectivity of such relations of satisfaction does not constitute in our sense an objective value."

[71] The affirmation of the judgment can be made by just reflecting on what one knows factually about the sport: If it is the case that switching to the inside edge on a lutz jump in the figure skating world is deemed a poor execution by the skater, then based on those measures of judging in the discourse of professional skating, the skater ought or should receive a penalty to his/her overall score. Though the "ought" in this judgment is not properly moral, the context by which the agent above is motivationally detached from his judgments about professional skating is not that dissimilar to an amoral agent like the Joker who presumably has moral knowledge of right and wrong within a societal context. The difference, though, with a character like The Joker is that his attitude towards morality seems more anti-moral than amoral.

agent's affirmation of X necessarily imbues the evaluation of X as normatively good or desirable. While Elizabeth Anscombe and David McNaughton both support this position, especially when considering the dynamics of John Milton's Satan in *Paradise Lost*, there were good reasons given for being skeptical of this view.

But consider the following amendment to the example above: Suppose the man happened to also be a judge in the competition. His motivational apathy was (mis)perceived by others as a rigid impartiality—a quality that is desired of referees in any sport. Similar to the situation above, he affirms the judgment that a penalty is appropriate after the skater's performance. Does the internalist's insincerity objection still apply? The man is no longer a passive observer in the stands but integrally involved in situation. An externalist may contend that it would be an open question whether his judgment was internally motivating or some other external factor influenced him to act on the judgment. The man's love for his mother or perhaps fear of the repercussions for nonconformity could be just as much a candidate for the source of his motivation rather than the judgment by itself. Does it necessarily follow that judgments provide an agent with the motivation to want to act in that way? The situation above seems to indicate the opposite, that motivation is not necessarily contained within the power of judgment, but that agent motivation can perhaps entirely depend upon some source external of a judgment.

In the realm of moral judgments, the motivation externalist is making a similar point. The amoral agent's disposition towards moral discourse is such that one has knowledge of right and wrong in an intellectual sense and is not moved by it. Unlike being conversant about professional figure skating, people utilize and appeal to moral language all the time. An amoral agent, unconcerned with questions of morality, comes into contact with moral language everywhere. As social animals, human beings are embroiled in the mores of their particular

communities.[72] There is hardly any escaping the moral assertions of others, even for amoral agents. So, it is not far-fetched—but actually quite plausible—for an amoral agent to score well on a moral competence or aptitude test. Ordinary moral agents, however, do not merely employ judgments but they actively immerse themselves in morality. The relation the mother has to professional figure skating is comparable to what moral practitioners generally have to moral discourse. They see themselves as moral agents and take their judgments seriously as action-guiding. This is not the case with how externalists depict the inner life of amoral agents.

To use Searle's terminology from his 1964 article "How to Derive 'Ought' from 'Is,' " there is an internal and external context to judgments involving "institutions."[73] The judgment "One ought to keep one's promises" is only relevant to agents who have opted—or have the desire to opt—into the moral institution of promise-keeping. Most agents normally participate in morality or have the desire to contribute to moral discourse. From their perspectives within the institution, to ask "Ought one to keep one's promises?" would rightly be an empty question.[74] The amoral agent, though, does not opt into this institution. Additionally, the agent is not internally motivated by the judgments made in moral discourse. This does not mean that amoral agents are ignorant about what moral judgments involve. They may know the intricacies of moral discourse better than some of the actual participants. While amoral agents are capable of mimicking moral judgments when they do not sincerely believe them to be true, it is open to debate whether this is attributable to amorality altogether or just particular instances of it. Unless amorality can be fundamentally linked with insincerity, it is difficult to conclude whether the judgments of amoral agents are what they really believe or judge to be the case.

[72] There can be applied to communities that utilize religious and political language as well.
[73] John Searle, "How to Derive 'Ought' from 'Is'," *The Philosophical Review* 73, no. 1 (1964): 51.
[74] Ibid.

There is a wealth of philosophical literature with reference to internalist and externalist debate on amoral agency.[75] The metaethical stances of both sides have far reaching implications that can enter into other areas such as epistemology and philosophy of mind (e.g. inquiries like "what is a belief?" or "what does it mean to believe something?"). To adequately explore every aspect of agent judgment and motivation would be too big a burden for this inquiry to handle. So the overall aim here is to focus on the objections most relevant to amorality and then later focus on evil—with a special emphasis on purely evil. Keeping with the benchmarks established in section 1.4, one should expect that internalism has explanations available to the alleged counterexamples above regularly offered by externalists or those sympathetic to a position contrary to internalism.

The most common internalist response to instances of amorality is to question whether or not such agents *sincerely* believe what they presumably judge (or knowingly affirmed) to be the moral rule. Similar to Ancombe's interpretation of Milton's Satan in section 1.2, such agents could merely be mimicking the normative evaluations of others in an inverted-commas sense. Returning to the judgment that "Being kind is a morally good thing to do", the amoral agent could really be saying "According to the moral standards of this community, being kind is a morally good thing to do". As such, judgment becomes a factual observation about other people's beliefs and not an expression of the agent's own beliefs. The amoral agent, like an anthropologist, is an outside observer that has an intellectual comprehension of the subject in question at the same level as those who have embraced or opted in the discourse, but is not motivationally invested towards acting on the knowledge. Similar to the ice skating example earlier an anthropologist can learn, take up the discourse, and by proxy be involved in the community. But none of this would indicate that the anthropologist believes (at least in a sincere, honest, way) any of the propositions of the discourse.

[75] For starters, check out McNaughton, *Moral Visions* (1984); Smith, *The Moral Problem* (1994); and Shafur-Landau, "A Defense of Motivational Externalism" (2000).

One could then interpret the externalist suggestions involving amoral agency as strange and simply untrue given that it is tantamount to the analogy that anthropologists, by becoming proficient in a particular discourse of a community, believe what they say in the same capacity as the practitioners themselves. Obviously there are professional standards of impartiality and resistance to assimilation which anthropologists in the field are subject to follow; but this is not the case with amoral agents. There is nothing inherent in moral discourse that obliges or compels such agents to demonstrate their beliefs or judgments. As such, it appears to be a working assumption that amoral agents are disingenuous about the nature of their own beliefs or judgments. On this point the anthropologist and the amoral agent both converge. While the externalist may have a point in suggesting that knowledge of something does not equate to having the motivation to act on it, this fact does not immediately strike against the MIT as it is currently understood. Internalists can preserve their account by couching amoral agents in terms of insincerity by not genuinely embracing a moral commitment to the discourse. Their expressed judgments have no inkling of intention and, thus, are nothing but a string of words that merely echo moral judgments found in discourse.

While all internalist accounts have access to this objection of amoral counterexamples, there are different accounts within motivation internalism and externalism that not only explain the relation between our thoughts and motivations but also touch upon moral psychology (e.g. good, evil, amoral agency). As much as motivation internalists draw on the standard reading of the MIT outlined in section 1.3, their accounts can vary depending on the degree or extent to which motivation enters into agent judgment. Some traditional or classical internalist models are rightfully dismissed by contemporary internalists because they easily fall victim to externalist counterexamples utilizing specific cases of amorality and/or akrasia.

One example is Platonic or Socratic internalism as it is conveyed by the paradox of the good in the dialogue *Meno* (among others). This paradox can be construed into two different

kinds: the moral and the prudential paradox of the good. The prudential paradox states that "all men desire good things" where good means "beneficial" or, inversely, those who pursue harmful things do so involuntarily.[76] The moral paradox is expressed later in the dialogue by Socrates' definition—virtue is knowledge—suggesting that all injustice or wrongdoing is done in ignorance.[77] While there is debate between scholars whether these two versions are independent or can be lumped together into one account, the implications of both paradoxes can impact the way internalism is formulated.

Regardless of any differences between the two, both versions of the paradox appear to deny the moral phenomenon of akrasia (i.e. weakness of will). The paradox of the good implies that no one would fail to be motivated towards the good if they had knowledge of the good. On a larger scale, it is part of a larger position known as psychological eudaimonism (PE). Rebecca Bensen-Cain, in *Socratic Method*, suggests that the paradox of the good—as it relates to PE—is something Socrates promotes as "true regardless of what the interlocutor claims to be the case because it belongs to human nature."[78] She defines psychological eudaimonism as the view that "all humans, by nature, desire the good where 'good' is understood as happiness or what conduces to happiness."[79] If virtue is knowledge, then any moral failure (including akrasia) is due to some underlying ignorance of the agent.

This rationale would also explain amoral agents in terms of insincerity, dishonesty, or perhaps even self-deception. Judgments from an amoral agent are not meaningfully expressed because there is a fundamental disconnect between one's desire for the good and belief of what is good. Thus, amorality is a product of ignorance in the sense that an agent's knowledge of the good would necessarily provoke the requisite action. Since amoral agents are defined as not

[76] Plato, "Meno", *Five Dialogues*, trans. G.M.A Grube (Indianapolis: Hackett Publishing, 2002), 66-67 [77b-78b].
[77] Ibid., [87-89].
[78] Rebecca Bensen-Cain, *Socratic Method*, (New York: Continuum International, 2007), 23.
[79] Ibid., 1.

being necessarily motivated by their moral judgments, then for Socrates they must not have knowledge of the good. If only amoral agents knew better in terms of acting on knowledge of the good, then they would be properly motivated to act.

This internalist model of moral motivation is flawed for several reasons. The first problem with this view concerns the appeal to "facts" about the nature of human motivation in order to explain the normativity of moral judgment. That is, as Bensen-Cain pointed out earlier, Socrates assumes rather than argues for the truth of psychological eudaimonism and subsequently the paradox of the good. Peter Railton suggests that the normative and motivational forces of morality do not seem so strictly bounded together—that perhaps externalists can better provide an *internal* story of how moral judgments relate to one's motivation to act.[80] It is doubtful to assert that human nature as a matter of fact is disposed towards the good. If one understood "good" as tautologically identical with whatever an agent regards as beneficial or desirable, then the Socratic internalist may have a case. But, as explored in section 1.2, one would be committing a hasty generalization and may omit certain irregular cases of moral agency (e.g. purely evil agency).

Socrates seems to suggest a psychological necessity between rational agency and human nature such that having knowledge of the good guarantees one will act in that manner. Just as the statement "All bachelors are unmarried men" embodies a strict conceptual definition between terms, the paradox of the good regards one's object of desire as a perceived or otherwise closeted good. The idea being expressed here can be reduced to the following: motivation is not and can never be inert whenever an agent consciously and correctly acknowledges what is good (and bad).

[80] Peter Railton, "Internalism for Externalists," *Philosophical Issues* 19, no. 1 (2009): 168.

This leads to the second and most obvious of problems with the Socratic view. Motivation *can* sometimes fail us, despite sincerely judging that a certain action is morally required. Also known as incontinence or akrasia, weakness of will is a moral problem that occurs when an overriding factor such as desire or some other conflicting circumstance generates inaction or hesitancy. In the dialogues *Protagoras* and *Republic*, Plato's Socrates addressed the potential impact that pleasure or desire can have on the one's rational deliberation. The case of Leontius' (sexual) "appetite" to look at corpses and his simultaneous revulsion towards this desire is remarked on in Bk.4 of the *Republic*.[81] This seems to indicate that Plato was at least aware of the potential overriding nature of competing desires.

But, for Plato, there is a notion that an agent is ignorant if desire can easily overturn one's knowledge—that the person did not really "know" in the first place. A person's correct knowledge of the good and regard for reason cannot be corrupted or overwhelmed by any desire no matter the circumstances. If such desires were to determine one's actions, then Leontius is nonetheless ignorant of what he knows or knew to be good. As such, Socrates' rather simplistic view of agency is that a person's actions ultimately demonstrate whether he/she has knowledge of the good or not. Amorality and akrasia are each the result of the agent's ignorance. The amoral agent not only fails to sincerely connected with their expressed moral judgments, but the agent also fails to even have knowledge of right and wrong in the first place. Similarly if akratic agents fail to act according to what they deem good, then the agents must not have really known. This internalist view is rightly disputed as it establishes an unwarranted guarantee that an agent's judgment will always succeed to motivate the action which adheres to the judgment. This internalist model of moral motivation comes off too strong in the determining power of motivation on pronounced judgment.

[81] Plato, *Republic*, trans. G.M.A Grube (Indianapolis/Cambridge: Hackett Publishing, 1992), 115-116 [440-440b].

To see where Socratic internalism may have gone wrong, one should re-examine the intellectualist[82] views of Plato's Socrates: PE and the paradox of the good. In terms of the city-soul analogy[83] which Plato outlined in detail throughout the *Republic*, there are several different motivational forces that pervade human nature. The rational nature of intellect, the spirited nature of emotions, and the appetitive nature of desires are things that can be found within every human being.[84] Socrates champions reason as the primary faculty that should rule the human being. Reason is superior to desire and should moderate its influence in everyday life. But the other two faculties, emotion and desire, often go hand in hand and do not aid the intellect's rational deliberations of good and bad. In fact, in the dialogue *Phaedo*, Socrates believes that to be a philosopher is to separate oneself from the attachments of the body—among them its emotional and appetitive urges.[85] Desires can impede even the most educated person. In these moments, ignorance (via emotion or desire) clouds one's reason and the "multiform beast….weaken[s] the human being within".[86]

The problem with Socratic internalism is the resultant view that if one truly had knowledge of the good, then one could not fail to act according to this knowledge. A human being's willpower that is aligned with reason and therefore with knowledge of what is good simply would not act otherwise. This seems to defy the reality of moral motivation as seen throughout moral discourse. People struggle all the time to act on what they believe to be good. Aristotle also echoes this concern about Socrates' rejection of akrasia in the *Nicomachean Ethics*:

> "[F]or it would be strange—so Socrates thought—if when knowledge was in a man something else could master it and drag it about like a slave. For Socrates

[82] Intellectualism is a position, of which Socrates represents an ancient Greek version, that embraces the intellect as the primary, superior, and most vital part of a human being. It is by the rational powers of the intellect that a human being should facilitate free choice and moderate one's conduct.
[83] See Plato, *Republic*, [368d-369b] for the initial introduction of the analogy.
[84] This is generally called Plato's tripartite view of the soul.
[85] Plato, "Phaedo", *Five Dialogues*, trans. G.M.A. Grube (Indianapolis/Cambridge: Hackett Publishing, 2002), 101-104 [64a-67e].
[86] Plato, *Republic*, [588e-589a].

> was entirely opposed to the view in question, holding that there is no such thing as [akrasia]; no one, he said, when he judges acts against what he judges best—people act so only by reason of ignorance. Now this view plainly contradicts the observed facts…"[87]

The "observed facts" in this case are the innumerable ethical situations in everyday life that can test one's resolve. Why does any agent sometimes hesitate with a decision, struggle with what he/she knows to be the morally right thing to do, or perhaps even behave/act contrary to their pronounced judgments? Genuine situations in ordinary life that follow these parameters can be easily identified. Common situational themes involve emotional influences such as addiction, infatuation, and perhaps even depression. For example most smokers today know the damaging effects of nicotine but are enthralled by its addictive influence. In cases of infatuation, an individual can be captivated by a damaging attraction towards some particular object of obsession to the extent that it diminishes other moral and non-moral concerns. With depression, it is not difficult to imagine a grieving widow struggling to fully maintain the motivating power of her moral beliefs (e.g. giving to charity, helping one's neighbor, etc) while in the grips of listlessness.

Competing motivations in these situations weaken their (moral) resolve and prevent the motivating power of judgments from coming to fruition. Either due to some competing motivation or motivation-sapping emotional influence, the smoker, the infatuated individual and grieving widow have their better motivational inclinations overridden and determined by a stronger influence that they cannot resist. This does not mean that these agents are unequivocally in a state of ignorance, but rather that the agent's self-knowledge of what is good was not strong enough to overcome other rival influences.

In addition to the above overriding conditions involving emotion, there are times when a situation can produce a clash between significant ethical values which may result in a

[87] Aristotle. "Nicomachean Ethics", in *Basic Works of Aristotle*, ed. Richard McKeon (New York: Modern Library, 2001), 1038 [1145b23-28].

kind of motivational paralysis. Such ethical dilemmas are generally a staple of any introductory ethics course in philosophy. One well-known example is Bernard Williams' Jim and Indians scenario that involves someone having to choose between killing a random person in order to save a large group of people or refusing and, as a result, having to watch helplessly as the whole group gets massacred. Williams points out that utilitarianism arrives at a seemingly obvious answer—to maximize the greatest good for the greatest number and save the group at the expense of one person—but fails to consider the cost of one's own personal feelings and/or integrity.[88] At times the morally right thing to do can not only be hard to perform, but can also come at a steep cost. A person's beliefs or personal judgments about killing do not relapse into a state of ignorance—as the Socratic internalist model seems to suggest—but rather the dire conditions of the situation overwhelm or override those concerns. Lack of (sufficient) concern for moral goodness, rather than a lack of moral understanding or knowledge, is the real problem here.

Akratic agents seem to know or judge X genuinely but lack the will to act on X due to some psychological hindrance. That is, on the level of forming moral judgments, the akratic agent correctly judges X but is obstructed by other factors. At face value, without some qualified revision(s) to the contrary, the Socratic internalist model suggests that akrasia is incoherent. But if moral failure can be genuinely caused by competing motivations (e.g. addiction, severe depression, etc), then this view is not adequately representing the judgment-motivation structure of human beings. If, as Stocker suggests, "the interrelations between motivation and evaluation are [both] various and complex"[89] then there must be more to human motivation than simply the state of possessing knowledge of what is good.

[88] Bernard Williams, "A Critique of Utilitarianism," in *Utilitarianism: For and Against*, (New York: Cambridge University Press, 1973), 98-99.
[89] Stocker, "Desiring the Bad: An Essay in Moral Psychology," 741.

This problem, though, seems to be unique to internalist formulations such as Plato's Socrates' intellectualism. Most internalists—and externalists for that matter—do not think that being motivated to do X necessarily *guarantees* X will happen. There is always the chance, as explained above, that a competing motivation or desire can override the normative force of a moral judgment and cause the agent to act in some other way. The pull of passion can oftentimes get the best of people's sincere judgments. By adopting broader versions of the MIT that allow for competing motivations to override an agent's better judgment, contemporary internalists can easily avoid the bulk of this criticism (when Socratic internalism cannot).

Externalists in the debate have the advantage of a very simple and comfortable position. In section 1.4, motivation externalism was referred to as merely the "skeptical negation" of the MIT. By not adhering to the necessity of an internal motivating element within judgment, externalists simply accept that the motivating reasons of certain agents can sometimes wildly deviate from their expressed judgments. The existence of amorality and akrasia is not problematic for the externalist.[90]

The internalist, however, does have the weight of initial plausibility on its side when gauging immediate intuitions about the matter. There is a general behavioral expectation that one's motivations to act are, *ceteris paribus*, causally linked to one's judgments. It would be odd, as mentioned previously, for one's mental faculties to reason and deliberate on some matter and then come to find one's motivational leanings are entirely removed from those thoughts. What would be the point of agent judgment in the first place if not to pinpoint one's motivational leanings?

[90] Samet-Porat ,"Satanic Motivations," 82, in particular makes this claim quite explicit. But that does not mean the motivation externalism is not without problems itself. A few will be mentioned in the next section.

These facets of an agent's will need not fundamentally conflict. The internalist and externalist can likely trivially agree about most "normal" cases adhering to Figure 1 outlined in section 1.3. As Peter Railton poignantly remarks, "every sane judgment-externalist will allow that moral judgments are so regularly accompanied by some sort of corresponding pro-attitude that we almost always feel that a special explanation is needed when someone who has made a seemingly sincere moral judgment appears entirely indifferent…"[91] Ultimately, what is being contested is the necessary and a priori relation between agent judgment and motivation—the core of the MIT. In other words, the concern is whether there are loopholes (i.e. counterexamples) in either metaethical reasoning and, if so, whether these cases can be reconciled in the same way contemporary internalists have moved past the Socratic understanding of moral motivation.

Supposing that Figure 1 in section 1.3 is the default framework of an agent's thought process, it does not directly confirm the MIT. Externalists only need a solid counterexample in order to defeat the necessary and a priori condition that judgments contain within themselves sufficient motivational force for an agent to *want* to act on it. If certain agents can genuinely approve of moral judgments and not be motivated to act on them, then an agent's judgment and motivation is at best only contingent; and as a result the MIT in its general form is flatly false—even if it turns out that only one particular sort of agency does not coincide with the internalist framework. Externalist tactics generally focus on critiquing internalism in this capacity because moral phenomenon like amorality and akrasia frequently occur in moral discourse. And what people observe of moral behavior in terms of pronounced judgments and their overall effect on motivation is sometimes contrary to a particular metaethical rule—in this case the MIT.

[91] Railton, "Internalism for Externalists," 168.

While one can view motivation externalism as a skeptical or negative hypothesis,[92] one could also simply consider externalism as an attempt to identify alternative mindsets relating to judgment that deviate from the MIT in general. Railton suggests this and further argues that motivation externalists themselves need to go internal, to supply alternative stories that draw a different parallel between moral judgment and motivation:

> "It seems to me that multiple 'motivational sets' could be consistent with sincere, full-fledged use of a normative kind term…As in natural kind language, so in normative kind language: *correct use* need not require a canonical idea or sentiment 'in the head'…"[93] (emphasis mine)

When motivation externalists formulate counterexamples to the MIT, what they are doing is attempting to describe an agent that correctly forms and applies moral judgments but at the same time those judgments are not inherently motivating in the way internalists would generally consider them to be. Externalists are taking the MIT, however it is presented, and investigating whether various conceptions of agency produce exceptions that do not quite adhere to this framework. As Wittgenstein once noted, "In philosophy one feels forced to look at concepts in a certain way. What I do is suggest, or even invent, other ways to look at it."[94] Externalists are, in some sense, Wittgensteinian. They are testing the parameters of internalism with possible counterexamples of agency that defy the metaethical rule. Given certain agent mindsets or thought processes, motivation may not be proof of taking a judgment normatively or vice versa.

Railton's sensible cad is someone who knows quite well the normative discourse of harassment and uses the term with the same normative force as other speakers, but he does not consider his judgments to be inherently motivating or action-guiding; in fact, he has no motivational impetus to act in any particular way despite his pronounced judgments

[92] See in particular Shafur-Landau, "A Defense of Motivational Externalism," 271.
[93] Railton, "Internalism for Externalists," 171.
[94] Ludwig Wittgenstein, "Lectures of 1946- 47," *Ludwig Wittgenstein: A Memoir*. ed. Norman Malcolm (Oxford: Oxford University Press, 1966), 43.

concerning the harassment of a fellow co-worker.[95] An internalist may suggest that such agents must have something wrong with them (e.g. insincerity, ignorance, etc) if their judgments lack motivational efficacy. But Railton does not think that Roger is simply roleplaying or mimicking his normative use of the term "harassment". Rather, he is using the shared meaning of the term without opting into the discourse himself (similar to the ice skating example described earlier). As such, Railton suggests that Roger and other agents like him can follow a path of *correct use* that is "parasitical"[96] on the shared meaning of a term within a community—those that constantly form judgments and appreciate their normative guidance first hand. In other words, Roger is committed "to using a normative concept with its ordinary, literal, shared meaning for a variety of reasons" but he is not "of one mind" with others in the discourse.[97] This example—as one may already notice—seems to relate in some ways to the free rider problem and its variations within ethics, social theory, political science, and economics.

The internalist may take issue with the parasitical nature of Roger's normative judgments, but as Railton is quick to point out "most of us stand in a similarly 'parasitical' relationship to the linguistic and scientific community when it comes to our use of proper names and natural kind concepts and terms, without involving the least insincerity or impropriety."[98] This puts the internalist in a tough dilemma. One cannot convict Roger of insincerity with his normative use of the term "harassment" without also convicting all other language users for similar offenses. If the internalist gives way to other paths of correct use that do not necessarily require motivation as proof of taking a concept normatively, then the necessary and a priori relation between judgment and motivation breaks down and the MIT weakens to the point that it becomes trivial. As quoted earlier from Railton, "every sane

[95] Railton, "Internalism for Externalists," 169-171.
[96] Ibid., 177.
[97] Ibid., 171.
[98] Ibid., 177.

judgment-externalist will allow that moral judgments are so regularly accompanied by some sort of corresponding pro-attitude…"[99] The flexibility of externalism to explain various accounts of agency is a strength that internalism sorely lacks in this matter.

Given Railton's descriptions of this case, Roger—a womanizer himself—understands the normative meaning of harassment but nonetheless correctly judges when others commit the offense. He is accused by others of being insincere, that he is "one to talk". Far from it, Roger actually cares about the accuracy of his judgments and "seems to understand the concept of harassment better than a number of his co-workers."[100] One presumably may have other problems with Railton's sensible cad, but the internalist cannot dismiss this agent's judgments in the same way as the amoral or akratic agent. Railton entertains the suggestion that Roger is employing a narrow, one-dimensional, normative use of harassment in which he escapes self-condemnation by exempting his own activities as not "textbook harassment."[101] Putting aside normative disputes on the meaning of harassment, ultimately for internalists they cannot say that Roger's lack of motivation is necessarily due to complications in his ability to judge the matter correctly or sincerely. The problem lies with his motivational compass, which internalists have said must be oriented towards wanting to act a certain way when it does not seem to be the case. Looking back at Figure 1 in section 1.3, Roger's case challenges the transition from S1 to S2. Perhaps Railton says this best in the following passage:

> "In practice, we ordinarily learn something about a person's state of mind when she makes a forceful normative recommendation or condemnation…The same is true, however, about ordinary belief: assertions normally convey information about the speaker's beliefs. But it would be an instance of Searle's 'speech-act fallacy' to attribute to an expression as its primary meaning a function the term characteristically, but not always or essentially, serve. Grasp of the way that facts about the speaker's state of mind are implicated by her judgment is evidently

[99] Ibid., 168.
[100] Ibid., 170.
[101] Ibid., 170.

> important for understanding all that goes on in ordinary normative discussions and exchanges of factual opinions. But none of this suggests that her state of mind must be part of the *content* of what she says…"[102]

The necessary and a priori relation between judgment and motivation is at the heart of the conflict here in Railton's case and also within the motivation internalism/externalism debate overall.

The example given by Railton above depicts an agent that does not quite fall into either amorality or akrasia, but rather stands as an in-between. And for that reason internalists can struggle to fit such agency within the MIT. Does the purely evil agent, who affirms evil as evil, also stand on its own conceptually? In the next section, purely evil agency will be included as an additional challenge (i.e. counterexample) to the MIT. Are internalists left without some flexibility of their own to explain agents that appear to be exceptions to the rule? This is still an open question, but the introduction of Kant's internalist model of moral motivation and subsequent account of evil in later sections may make headway towards answering it. Similar to amorality and akrasia, the motivations that inspire and structurally underlie evildoings are not small matters for discussion.

[102] Ibid., 177.

2.2 Evil Agency

How do internalists explain the motivations of evil agents? The conceptual and psychological treatment of evil in chapter one was the prelude to a larger scale analysis of motivation internalist and externalist accounts of evil. In particular, there is a conception of pure evil agency that seems to turn the MIT on its head—more so than any imagined case of amoral agency examined in the previous section. Just like how an amoral agent can score well on a moral competence or aptitude test and at the same time have such knowledge fail to be entirely motivating in practice, the purely evil agent has a strong intellectual grasp of morality but is motivated to do what is considered evil rather than good. For purely evil agents, the judgment that "X is good" represents motivation to do the contrary. As discussed in section 1.1, purely evil agency is a rarified case that hardly plays an actual role outside of the literary and cinematic realms. Nevertheless, its rarity should not discount any explanation (or lack of understanding) on the conceptual and psychological level. One should remain open to varied rationalizations of pure evil from the motivation internalist and externalist positions.

To briefly summarize, the difference between perverse and purely evil agency comes down to whether one performs such acts instrumentally or intrinsically. A perversely evil agent is a corrupted agent that performs actions with some particular end in mind other than for the act itself. There is an agenda or overarching motive that fuels the agent's drive to do evil acts. Hence, a perversely agent looks upon his/her actions instrumentally—as a means to an end. Wealth, power, self-righteousness are some (of presumably many) goals which cause one to commit evil acts as a means.

Purely evil agents, on the other hand, operate within the same moral context as ordinary moral agents but have an inverted motivational attitude that is attracted to evil rather than good.[103] Section 1.2 addressed the intricacies of perverse and pure evil motivation, ultimately closing with the suggestion that purely evil agents do not operate under a twisted conception of good when embracing the label of "evil". Personalities like Adolf Hitler—though his actions were certainly monstrously evil—are unambiguously perversely evil. Hitler did not commit genocide for the sake of genocide. The act of genocide was not itself treasured as an end but as a means. He did not bring destruction and death to millions for the sake of just doing it. By contrast, St. Augustine of Hippo seems to describe the purely principled stance that embodies pure evil when recalling stealing pears as a child:

> "I stole something which I had in plenty and of much better quality. My desire was to enjoy not what I sought by stealing but merely the excitement of thieving and the doing of what was wrong…Even if we ate a few, nevertheless our pleasure lay in doing what was not allowed…I had no motive for my wickedness except wickedness itself. It was foul and I loved it."[104]

Augustine's childhood theft mirrors in many ways the phenomenon of "thrill killing" that was briefly mentioned in section 1.1. While purely evil agents may experience some form of instrumental pleasure from their actions, the pleasure or other benefits of the evil act are secondary to the primary motivation: to simply do it is reason enough. Evil for evil's sake requires a character with principle and resolve to the same degree as one who is "good for goodness sake."

If this explanation is not enough, there is another contrast to be made between perverse and pure evil. In the case of Milton's Satan which is outlined in section 1.2, it seems that not all evil agents express their evil intentions under the same context. When Satan uttered "Evil, be

---

[103] Samet-Porat, "Satanic Motivations" (2009), calls such agents "satanic" and Immanuel Kant, in *Religion within the Bounds of Bare Reason*, calls them "diabolical" or devilish. Whatever term one uses is ultimately a matter of personal preference. This inquiry will stick with the label "purely evil".

[104] Augustine of Hippo, "The Depth of Vice: from *Confessions*", trans. John K. Ryan. *Vice and Virtue in Everyday Life* 8th. ed. Christina Hoff Sommers|Fred Sommers(Belmont: Wadsworth, 2010), 336.

thou my Good" in *Paradise Lost*, his understanding of good and evil is starkly different from many others that fall under the label "evil". As suggested in 1.2, Satan's utterance can be interpreted as using a formal sense of "good" rather than an evaluative one. These different functions reveal a conceptual divide among some instances of evil-doing. Most, if not all, agents make extensive use of the terms "good" and "evil"; but perversely evil agents evoke those terms within a revised, twisted, moral framework of their own devising. As such, they form moral judgments under a *de re* context—Latin for "of the thing". This sort of understanding picks out the particular thing which the terms are supposed to represent. The perversely evil agent subsumes a particular thing or value (e.g. power, wealth, self-righteousness, etc.) under the label of "good" and, by doing so, operates under a twisted moral code in direct opposition to the initial use of the term. Hitler's vilification of the Jews as evil, for example, reoriented his moral value of good to the extent that genocide could be justified in his view. In other words, the perversely evil agent does not value evil as evil but rather evil as power, evil as happiness, or some other perceived good.

A *de re* understanding associates the term "good" with a particular thing an agent identifies, which allowed Hitler to "justify" genocide in accordance with his twisted agenda. The purely evil agent, though, does not redefine what is good or evil. Rather, the agent works within ordinary moral discourse and chooses evil as it already stands. This understanding operates under a *de dicto* context—Latin for "of the word". The purely evil agent relates to good and evil according to the meaning of the words themselves, whereas a perversely evil agent dictates good and evil according to whatever specified interest with which they identify. The purely evil agent's aims just happen to lead to the attainment of power, wealth, or some other "good" but the perversely evil agent aims for those things at the start. Purely evil agents do not seek to justify themselves as good or right. Doing evil is its own reward.

As with amoral agency, there are various ways for an internalist to give account of evil. Plato's Socrates explained ordinary (i.e. perverse) evil as a fundamental mistake of one's knowledge of what is good.[105] Evil is the result of ignorance. Hitler presumably did not think his actions were bad. He believed what he was doing was good or morally justifiable. Thus, one could say in the Socratic vein that Hitler was blinded by an overall ignorance of the actual good. If he had known better, Hitler would have not done those horrific actions. While ignorance can explain the motivational failure of some evil agents, there are other instances of evil that do not seem to fit Socrates' rationale.

In section 2.1, it was suggested that Plato's internalism seems to place too strong a guarantee on the success of an agent's judgment for inducing a motivation to act. It is likely that Plato's Socrates would have also considered the concept of pure evil fundamentally incoherent given the assumptions of the Socratic paradox of the good. Like some cases of amorality and akrasia, purely evil agents cannot be straightforwardly accused of ignorance because these agents by hypothesis speculatively know the difference between good and evil. To complicate the matter further, purely evil agents are motivated to do evil because of the fact that they judge it to not be good. This is presumably a contradictory notion for Plato's Socrates, akin to the concept of a married bachelor or a four-sided triangle.

The explanation that pure evil is merely a closeted perverse evil was questioned in section 1.2. If pure evil is conceptually plausible, then Plato's Socrates (or any other internalist account of evil for that matter) cannot relegate the purely evil agent to a closet lover of what is good without consequence. Plato, as evidenced in the *Republic*, recognized that the power of desire was irrational and needed to be controlled by the rational powers of human agency. The desire to do evil for evil's sake should be no different for Socrates; purely evil motivation is a mark of irrationality. This internalist explanation of motivational failure is what Andrew

[105] See in particular Plato's *Meno*, 66-67 [77c-78b].

Sneddon refers to as "Rational Moral Judgment Internalism" or rational motivation internalism.[106] While Plato's internalism fails for other reasons, there are many other internalist accounts of evil that follow a similar rationale.[107]

In lieu of the objections centered on Socratic internalism, contemporary internalists can simply claim instead that there is sufficient motivation in one's judgment to *want* to be motivated to act given no contrary overriding influences. Though the strength of the relation between motivation and judgment is not strictly binding, internalists can still maintain an agent's sincere judgment that "X is good" and motivation to act on that judgment as necessarily internally connected—one simply *must* imply the other. There is an additional explanatory model that internalists can co-opt with the rationality conditions above, which Sneddon refers to as "Psychological Moral Judgment Internalism."[108] This kind of account attributes motivational failure to agents due to immaturity, deficiency (in judgment), or mental illness. The measurement here is in terms of psychological stability or "normalcy."[109] Akratic, amoral, and evil agents in general are ignorant by varying degrees of immaturity or psychological deficiency. Inexperience or mental depravity (i.e. corruption) lends an agent to be deluded and wrongly choose what he/she judges good and bad. The extent of this self-delusion is presumably rather extreme for the purely evil agent.

Psychological moral judgment internalists, however, are subject to externalist criticisms as well. Whereas rational moral judgment internalists appeal to the irrationality of an agent in order to explain the disconnection between moral judgment and motivation, psychological moral judgment internalists appeal to mental discrepancies that cause agents to purposefully form bad

[106] Sneddon, "Alternative Motivation," 43.
[107] One example is Michael Smith's *The Moral Problem* (1994) which utilizes a model of "ideal rational agency" to defend motivation internalism.
[108] Sneddon, "Alternative Motivation," 43.
[109] Assuming that there are standard psychological markers for determining normal and aberrant behavior, evil agents can be construed as anomalous deviations resulting from psychological defect. However, this suggestion is not without potential objections.

judgments. Both positions highlight the same overarching claim that the agent does not really make a sincere moral judgment, but explain it using different methods.

Psychologically, the internalist could argue that evil agents suffer from warped (or otherwise ineffective) upbringing via mental defect or abuse. Aristotle, at the beginning of his *Nicomachean Ethics*, suggested that his ethical system will not profit a student if he/she were "inexperienced in the actions that occur in life" and lacked the proper development of character.[110] In other words, moral knowledge is useless if the agent did not already develop a stable moral character on which to appreciate moral improvement. If an agent, similar to the profiles of many serial killers, grew up amid abuse and negligent moral guidance then it is highly likely that any moral education in later years would not be profitable.

But purely evil agents, as they have been construed in the previous chapter, are just as morally educated and situated to live a moral life as any ordinary agent. Thus, like the irrationality objection above, any psychological rationalization of pure evil seems to miss the mark and instead picks out some perceived manifestation of perverse evil. Samet-Porat uses the label of "preferential evil" to describe the type of mindset of those who do a morally wrong action because "[the agent] prefers some other end to the avoidance of moral wrongdoing."[111] However, preferential evil is still a step away from pure evil because "the depravity is not a by-product…of that state of affairs."[112] A purely evil agent does what is morally wrong precisely because he/she judges it to be the morally wrong action.

Purely evil agents are a rather daunting case for internalists to explain, even within both rational and psychological explanatory models of internalism. This is because such agents seem to be different than other instantiations of evil; they are not twisted, ignorant, or clearly irrational

[110] Aristotle, "Nichomachean Ethics," 936-937 [1095a1-10].
[111] Samet-Porat, "Satanic Motivations," 78.
[112] Ibid.

in the same way as perversely evil agents nor are they motivated by some other end than moral wrongdoing as are preferential evil-doers. Further, purely evil agents do not seem to suffer any mental defects or immaturity other than choosing to be evil (if such a thing can be considered a defect). There is no reason to think that the intellectual faculties of purely evil agents are less acute than their (morally) good counterparts. Simply, the agent under seemingly normal conditions affirms what is bad or evil on principle.

Lockie describes the internalist process of explaining this sort of evil as an instance of inverted internalism:[113] a case where internalism is turned on its head through a negatively confirming instance. Unlike the amoral agent who lacks any motivation at all towards moral judgments and the akratic agent who is temporarily separated from the influence of their judgments, the purely evil agent "has the cognitive ability to discriminate immorality correctly enough, yet is held to be dissociated from the normal response to that immorality by being attracted to evil in itself."[114] The moral judgments and motivations in this case are disassociated in that one can judge a thing good but not just be entirely unmoved by that evaluation. The purely evil agent's evaluative judgment produces the motivation to do the contrary. It may be tempting for the internalist to reject this sort of agency out of hand as yet another mask of perverse evil. But there are some that are hesitant to dismiss this conceptualization.

An externalist insists that the evil agent could rightly and sincerely judge what is good and not be otherwise motivated to act in the direction of their judgment. By disassociating the necessary connection between expressing a moral judgment and being motivated to act on it, the externalist does not have difficulty with explaining abnormal types of agency (e.g. amoral, akratic, evil). Whereas the internalist posits motivating power internal to the judgment itself, externalists do not readily place the agent's motivating power to act within the confines of their

[113] Lockie, "What's Wrong with Moral Internalism," 19.
[114] Ibid., 19-20.

judgments. That is, the judgment alone may be just one of several external factors that can bring about the motivation to act. Many externalists suggest that a corresponding desire or sentiment is needed to accompany an evaluative judgment and that perhaps the sentiment, not the judgment, contains the motivating influence.[115] The corresponding desire does not seem to be explicitly internal to the judgment but an external feature of the agent's mental state and circumstances. Thus, as an alternative account to the internalist framework outlined in section 1.3, the motivation to act may not be necessarily tied to the normative qualities of agent judgment.

Sentimentalism or moral sense theory reflects the view above that a moral belief or judgment is ultimately grounded in sentiment or emotion.[116] Having knowledge of what is good does not contain motivating power, but rather it is one's own individual makeup and emotional development that determines how moral judgments affect one's conduct. Even if the moral judgment is legitimate (i.e. sincere), it does not necessarily inspire an agent to want to act on it. Without the power of sentiment in human beings, moral judgments by themselves are inert. As a result, the externalist can attribute to evil agents discernable knowledge of what is good; yet their motivations to some external sentiment makes that knowledge just a series of claims among many others.[117]

Samet-Porat, however, accuses the externalist approach of being too simple and perpetuating a counter-intuitive psychology.[118] If evil agents stand on the same epistemological and psychological footing as ordinary agents then externalism (at least when considering an externalist position that adopts a Humean sentimentalist understanding of moral motivation)

[115] Shafer-Landau (2000) and Zangwill (2003) in particular make suggestions along these lines.
[116] Philosophers that traditionally represent this view are David Hume and Adam Smith.
[117] If, like the emotivists, one collapses moral belief into purely expressions of emotional approval or disapproval, then there would be a case of motivation internalism and not externalism. Since the motivating power of judgment and the judgment itself are located in the same thing via a non-cognitive emotive state, some forms of sentimentalism may be more suited toward internalism. But this is not the case for all forms of sentimentalist theory. By contrast, cognitivism and sentimentalism together make an intuitive case for externalism.
[118] Samet-Porat, "Satanic Motivations," 83.

makes reason impotent and "devoid of any motivational efficacy."[119] While externalists may have a point in criticizing the rigidity of the necessary and a priori relation between agent judgment and motivation, they also need to avoid the conclusion that there is no relation altogether and agents are subject to mere motivational whim.

In other words, as Samet-Porat's objection above suggests, the externalist view gives too much latitude to evil agency to the extent that an agent's motivations can be inclined towards evil (in itself) as much as it can be inclined towards the good. Most people want their metaethical beliefs to complement their normative deliberations. That is, an explanation of evil should also be an indictment of the underlying problem(s) of being evil. While externalists have the advantage of an easier, more flexible view of moral motivation, they are in danger of undermining what Christine Korsgaard calls "the normative question"[120]—why should I be moral (in the first place)? At bottom, externalists (at least those dependent upon the Humean theory of moral motivation) suggest that "no desire is contrary to reason… [thus] desiring what is bad *qua* bad is not irrational and therefore raises no special problem."[121] This is quite a claim for anyone in moral discourse to integrate into their normative ethical theorizing. Due to this factor Samet-Porat believes, contrary to Shafur-Landau's view mentioned previously, that externalism has the burden to clearly refute the MIT.[122]

As previously outlined in section 1.4, burden of proof is not the focus of this inquiry but rather how both positions in light of these accusations fortify their explanatory rationale on evil. Externalists need to maintain a balance between the rejection that judgments are necessarily internally motivating and lapsing into arbitrariness and no mitigated structure of motivation altogether. For internalists, though, one motivating sentiment cannot be just as good as another or

[119] Ibid., 83.
[120] Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), 9-10.
[121] Samet-Porat, "Satanic Motivations," 83.
[122] Samet-Porat, "Satanic Motivations," 83.

else one would have to conclude that moral goodness is no less preferable than wickedness. If it happens to be the case that agents can correctly (and sincerely) judge things to be the case and not have any motivation to want such things, then moral goodness is put on the same level—in terms of motivational preference(s)—as its contrary.

The rationality and psychological conditions of internalist explanations also serve to critically exhibit the weaknesses of abnormal types of agency (e.g. amoral, akratic, evil). The reason that those agents seem to be exceptions to the MIT is because of a certain failing or weakness within their judgment or capacity to be motivated. As such, one may suggest that the fault lies not with the MIT but rather with the agent in question. Amoral agents seem to correctly judge what is good but their sincerity is called into question. After all, how can one possibly embrace a judgment and not feel any (motivating) incentive to act on it? The judgment itself, for internalists, does not appear to be a genuine expression of the agent; it sounds like a moral judgment but there is no further indication that the agent "owns" the judgment except by speech alone. For the internalist, this must and should be scrutinized. Similarly, akratic agents are compromised by some overriding factor of emotion. The motivation for these agents to want to act on their judgments is present, but something else has prevented the wanting from being actualized (e.g. depression, passion, trauma, etc.). If Railton's case above is able to penetrate the MIT, then internalists can simply relegate it as an isolated anomaly. As Simon Blackburn once noted, "externalists can have individual cases, but internalism wins the war."[123]

Thus, there is an impasse between both metaethical positions and, like the theist and atheist debate in philosophy of religion, there seems to be no reconcilable conclusion in view—just two views that necessarily excluded the other. This disparity seems to also be the case when it comes to explaining evil. Internalists and externalists take their accounts into polar opposite

[123] Simon Blackburn, *Ruling Passions: A Theory of Practical Reason* (Oxford: Oxford University Press, 1998), 45-46.

directions. The internalist view is inclined to construe the evil agent (depending on the case) as an outright ignoramus, a closet lover of the good, a socially or psychologically unstable misfit, a deceiving or self-deceiving free-rider, or simply as an irrational sensualist completely succumbed to selfish desires. The externalist view, through the use of counterexamples, tries to establish a contingent relationship between the sincere expression of a judgment and its internally motivating qualities. As such, externalists are attempting to secure other (external) sources of motivation for an agent's judgment to make outlier cases involving amorality and purely evil agency more intelligible. As Blackburn hinted above though, this already gives up part of the game to internalism. One could go so far as to say that externalism itself is parasitical and depends upon internalism when it comes to giving an account of normal, everyday cases of moral judgment. Railton and presumably other externalists as well recognize the intuitive appeal that the MIT offers under normal circumstances.

Returning to purely evil agency, internalists should be cognizant of one particular problem. While rational and psychological parameters might describe and explain most evil-doers, pure evil agency still appears to be an exception and rises above these rationalizations. The explanations that internalists generally attribute to amoral and akratic agencies do not translate well with purely evil agency. Unlike perversely evil agents such as *Star Wars*' Anakin Skywalker, purely evil agents are not psychologically compromised. Nor are they ignorantly pursuing some twisted perceived good as Hitler or psycho-sociopathic criminal like Ted Bundy. There is a remarkable kind of self-awareness within the purely evil mindset. Such acute reflection can take internalists by surprise because it is the agent's knowledge of what is good and evil that is used against them. Samet-Porat similarly points out that "[purely evil] individuals seem to cut through the internalist theory. They preserve the structure of practical rationality while inverting the direction of motivation, so that a course of action is attractive as a result of a

judgment that it is bad."[124] Regardless of any substantial weaknesses on the part of externalism, internalist must have some answer for this sort of agency.

With some preliminary background of the debate between motivation internalists and externalists established as well as how both positions can begin to explain evil agency, the next section will take up this problem caused by purely evil agency and will be further explored in the third chapter of this inquiry. Immanuel Kant's internalism rejects the possibility of purely evil agency, at least in any human form. This position will be outlined in section 2.3 and further explored in relation to the previous chapters of this inquiry.

[124] Samet-Porat, "Satanic Motivations," 79.

2.3 The Kantian Approach

Examples involving amoral and akratic agency have dominated the debate between motivation internalism and externalism. The previous sections have explored some of the strongest internalists explanations for these counterexamples. But pure evil agency seems to rebuke all those explanations. Some have attempted to associate pure evil as a sophisticated, perverse evil. Others have rejected the conceptualization out of hand as impossible. Immanuel Kant, in his seminal work *Religion within the Bounds of Bare Reason* (referred to henceforth as simply *Religion*), seems to embrace the latter approach when it comes to human agency. Kant brings a whole new analysis of pure evil as humanly impossible that also fits comfortably with the MIT described earlier in section 1.3. This section will give a brief outline of the basic foundations that compose Kant's internalism. Even though there are religious overtones in his work,[125] this should not detract from the metaethical insights that Kant's work can bring to this inquiry.[126]

Immanuel Kant presents a rather sophisticated internalist explanation for evil. With his deontological ethics firmly in the background, Kant approaches evil from the standpoint that morality is eminently rational.[127] To be moral is to exercise one's capacity as a rational being. For an agent to act contrary to morality, one makes a choice that is not supported by the power of reason alone. Naturally, the greatest obstacle that interferes with the motivating power of

[125] Stephen Palmquist's Introduction to *Religion* tries fairly to curtail the tendency by antireligious scholars and readers alike from misinterpreting Kant's aim(s) in his work and the extent of his religiosity (if any). It is a matter of speculation as to Kant's own private religious beliefs in his adult life; and there is no easy way to translate Kant's personal leanings into a modern religious context.

[126] Pablo Muchnik, *Kant's Theory of Evil*, (Lantham: Lexington Books, 2009), 3.
Similarly, Pablo Muchniks' chapter "On the Alledged Vacuity of Kant's Concept of Evil" attempts to mitigate what Goethe and others considered as a "'stain' in the mantle of [Kant's] critical philosophy"— that Kant was simply perpetuating the same religious prejudices of his time. Far from it, Kant's work in *Religion* serves to complement and continue the inquiry from his three *Critiques*. Whatever support religious followers may glean from Kant's work is their prerogative. Kant's *Religion* is a work of philosophical theology rather than religious apologetics.

[127] There are some significant conceptual considerations that Kant includes in his account of evil and they will be explored in chapter three. Further, this chapter's earlier outline of amorality and akrasia will be complemented by Kant's three grades of evil in section 3.1.

morality is the influence of sentiment (e.g. the desires or passions). As a result human beings, as rational and sentimental beings, are constantly having their beliefs, judgments, choices, and ultimately actions torn between two motivational forces. Kant's framework for defending this view is based on largely a priori reasoning. It is not enough to make observations or introspect on the process by which judgments are internally motivating. Empirical analyses of moral judgments merely scratch the surface of the motivational depths of agency. For Kant, one must examine the will and the process by which an agent cultivates their moral disposition underlying maxim formulation.

Even though Kant defines the will as "a kind of causality belonging to living beings so far as they are rational,"[128] the term is employed in the context of two distinct functions—*Wille* and *Willkur*. Henry Allison, in *Kant's Theory of Freedom*, outlines a careful analysis distinguishing the two senses of the term.[129] "Wille" represents the legislative function(s) of the will—the capacity of an agent to reason practically and form maxims. "Willkur", on the other hand, represents the executive function(s) of the will—the capacity of the agent to choose maxims such that they conform (or fail to conform) to the moral law. Both together constitute the foundations of Kant's theory of freedom and subsequently his theory of moral motivation.

The will, both legislatively and executively, is an important starting point and deviation from previous internalist positions because Kant places it as the motivational origin of one's moral judgments. Rather than purely from a state of knowledge (or lack thereof), agent motivation and ultimately one's moral status as good or evil is situated in the will. In *Religion*, Kant identifies the source of evil as heterogeneity within the human will—the "power of choice itself" to subjectively determine maxims that are contrary to the moral law.[130] It serves as the

[128] Immanuel Kant, *Groundwork of the Metaphysics of Morals*, [AK 4: 446].
[129] Henry Allison, *Kant's Theory of Freedom*, (Cambridge: Cambridge University Press,1990), 129-136.
[130] Immanuel Kant, *Religion within the Bounds of Bare Reason*, trans. Werner S. Pluhar (Indianapolis: Hackett Publishing, 2009), 21 [AK 6:21].

ground of the agent's incentives to act in one way as opposed to another (i.e. respect for the moral law or indulgence in one's desires). Kantian literature credits Henry Allison dubbing this view "the Incorporation Thesis".[131] An agent's adoption of a motivating incentive for action is to have it "incorporated" into their will. Thus, by associating agent choice with what one (motivationally) prioritizes in action, Kant's internalism fuses agent judgment with necessary or obligatory motivating force as a synthetic *a priori* truth.

The Incorporation Thesis sets up the process by which a human will is either good or evil. One's actions are not properly evil; rather it is the principle being chosen and affirmed within the action that is the source of evil. In other words, actions by themselves are not sufficiently able to indicate the moral status of the agent in question because circumstances (among other things) can make an action seem like something it is not. Such theorizing is common, especially in consequentialist ethics. But Kant regards this thinking as akin to anthropology; such "principles of mere experience" are inferior to the grounding of a pure morality based on the a priori concepts of pure reason.[132] It is the rational capacity to choose "in which the agent must invest the moral incentive with motivating power" that indicates one's moral status as a good or evil will.[133] Thus, the agent must put one's incentives toward particular maxims into the proper order of priority. What the agent does in action is largely irrelevant to the constitution of that agent's will. In fact, how an agent acts in the first place is utterly dependent on what the agent gives priority to in their incentives.

While there are specifics to Kant's normative ethics in terms of the criteria maxims must meet in order to be considered properly moral (via the three formulations of the categorical imperative), Kant's proposed decision procedure is itself utterly formulaic. The categorical imperative outlines a thought process by which an agent can rationally determine maxims that

[131] Allison, *Kant's Theory of Freedom*, 5;51.
[132] Kant, *Groundwork*, [AK 4:38– 390].
[133] Muchnik, *Kant's Theory of Evil*, 11.

adhere to the moral law. Metaethically, this represents the first order stage of agent deliberation. There is a second order stage required of the agent before the categorical imperative can be appropriately used. One must be oriented towards having a good will and the incentives that support it. That is, the agent must "take an interest in the good" and have their priorities directed towards a rational duty to respect the moral law.[134] Pablo Muchnik calls this the "ultimate principle of maxim-selection" or meta-maxim, in Kant's German *Gesinnung* (disposition).[135] As such, the agent incorporates into their will whichever incentive primarily motivates. By virtue of making a judgment, the agent had at some point prior to the judgment made a fundamental choice about which motivating incentives will take priority over others.

Kant's rigorism requires the moral status of agents to be interpreted in only two distinct ways, as either morally good or morally evil. While experience may seem to support the contrary, there cannot be any ambiguity when it comes to the Gesinnung of an agent or else "all maxims run the risk of losing their determinateness and stability."[136] To be clear, there can be plenty of evidence that moral actions and perhaps much of normative theorizing in general cannot fully permit rigorism. But, for Kant, this cannot extend to the basis of the agent's principle of maxim-selection. In this case it simply has to be one of two options. Either an agent prioritizes the incentives of the moral law or some other inclination which Kant broadly refers to as "self-love".

In Kant's *Religion*, the primary distinction between having a good will or an evil will is based entirely on one's Gesinnung in terms of one's chosen order of priorities. A good will possesses the motive to perform the moral law for itself—to act purely from a rational duty to respect the moral law. On the other hand, the evil will chooses to prioritize inclinations generated by the principal of self-love. Pablo Muchnik illustrates the difference as similar to Kant's

[134] Ibid., 12.
[135] Ibid., 45.
[136] Kant, *Religion*, 22 [AK 6:22].

distinction between categorical and hypothetical imperatives (as well as autonomy/heteronomy). The hypothetical imperative "If you want to be trusted by others, then you ought not to steal" is vastly different than the categorical imperative "You ought not to steal!" To have a will that prioritizes the incentives of self-love is to ground one's maxims on the condition of a desire. The person's choice to follow the former imperative rather than the latter situates their will on the determination of some other factor or circumstance. As such, the agent has adopted the criterion of heteronomy instead of autonomy via the self-sustaining, rational principle in the moral law.

Ultimately, for Kant, these commands fail to be morally substantive because "the conditional character…forces reason to look for further conditions to ground and justify them" culminating into an infinite regress.[137] The problem is at the very roots of the agent's Gesinnung. The incentives of self-love lack justification because "[t]hey require the agent to will something because something else is willed, giving the pathological interest dominance over the practical one."[138] The moral law, however, is unconditioned and satisfies reason's demand to be acceptable for every rational being.[139] None of this necessarily entails that the evil Gesinnung would perpetuate actions deemed purely or even perversely evil. The person may or may not steal in this case or perhaps never at all. Whether or not one can associate or demonstrate some evil action with agency is irrelevant to agent Gesinnung. Regardless of how much an agent's actions are morally praiseworthy (or not), the incorporation of self-love as one's principle of maxim-selection is the source of evil. By giving in to incentives not conducive to respecting the moral law, agents have put themselves in the position of committing evil actions. Moral goodness and autonomy both seem to have an a priori connection which the incentives contrary to the moral law cannot access.

[137] Muchnik, *Kant's Theory of Evil*, 30.
[138] Ibid.
[139] Ibid.

Agent volition (i.e. free will) is an important aspect of Kant's internalism when considering the relations between morality and the incentives of rationality and sensuality. Kant identifies the human will as structurally heterogeneous which is to say that human volition is determined by both objective conditions (e.g. reason and incentive(s) of the moral law) and subjective conditions (e.g. passion and the incentives of self-love). Humans have freedom beyond that of an animal's habitual will or a divine, purely rational will.[140] Rationality allows agents such as human beings the ability to act against entreating desires, despite their persistence, and similarly sensuality allows agents the free choice to not be determined by the dictates of reason. Whereas (most) animals compulsively act on their desires, human beings—agents with a rational and sensuous nature—are not at the mercy of their desires or reason and can subjectively determine their own actions. Neither reason nor the sensuous impulses determine the human will unless the individual deems it an incentive.

It is important to note how autonomy relates to morality and particularly one's standing as agents. Kant's conception of freedom does not equate to an unfounded lawlessness. That is, freedom does not operate in a vacuum absent of any configuration. A lawless free will would be an absurdity and counteracts the rational nature humans inherently possess. Autonomy, like reason, must follow a structure in order to clearly and consistently make sense of free choice. Allison and other scholars refer to this critical insight as Kant's "Reciprocity Thesis" since the claim here is that morality and freedom are reciprocal concepts.[141] That is, one's commitment to the moral law is an expression, one and the same, of autonomy:

> "Hence, freedom of will, although it is not the property of conforming to laws of nature, is not for this reason lawless: it must rather be a causality conforming to immutable laws though of a special kind; for otherwise a free will would be self-

[140] Ibid., 8-9.
Pablo Muchnik says further that "the depth of freedom belongs to the human in-between." The heterogeneous structure of the human's will to choose between the self and the moral law bestows the kind of freedom inaccessible to the animal or holy will.
[141] Allison, *Kant's Theory of Freedom*, 201-202.

> contradictory…This is precisely the formula of the categorical imperative and the principle of morality. Thus a free will and a will under moral laws are one and the same."[142]

Just as the requirements for morality arise from rational agency, Kant asserts a fundamental entailment between freedom and morality. The heterogeneous nature of the human will allows for individuals the choice to prioritize incentives towards the cultivation of a good or evil Gesinnung. And one's Gesinnung can presumably change if the human will incorporates the opposite principle of maxim-selection.

For Kant, all of these a priori steps are the basic building blocks of his internalist view. It allows him to ascribe to all agents—to every human being—the propensity to evil by virtue of possessing a rational nature and at the same time sensuous impulses which can potentially overturn the incentive to respect (i.e. prioritize) the moral law. One's Gesinnung is the basis by which each agent has decided how they will select maxims that will be action-guiding. An agent develops an evil Gesinnung when one incorporates the principle of self-love, which is simply to say that the agent is subjectively determining his/her maxims by incentives not solely or directly concerned with the moral law. Yet human beings also have a predisposition towards the moral law in which the moral law would be completely irresistible if only "no other incentive acted against it."[143]

Thus, rather than being a particular mindset or approach, evil seems to be a *condition* that all agents share as a burdensome requirement for the kind of freedom animal and divine entities could never obtain. In *Religion*, Kant affirms that human beings—by virtue of their heteronomous nature—have a universal propensity to evil. Furthermore, Kant strikes at the heart of what this inquiry is searching for by immediately stripping bare the physical appearance of evil

[142] Kant, *Groundwork*, [AK 4:446-447].
[143] Kant, *Religion*, 40 [AK 6:36].

to its core feature(s). This is precisely the approach needed to procure some answers about the inner workings of evil.

To begin, Kantian internalism establishes agent motivation through an a priori relation between the concepts of pure reason and human beings' predisposition to the moral law. As Allison explains, "…the mere consciousness of the law of itself produces something like a 'proattitude' toward the law in the agent, which in turn constitutes the conative factor in action from duty."[144] If rationality and morality are intimately connected, the moral law has its own (sufficient) motivational power when an agent genuinely formulates moral judgments. Naturally there are overriding considerations that can result when agents are seduced by the incentives of self-love over the moral law.[145] But, as previously explained, the actions produced by self-love do not constitute what is principally evil; rather, the agent's legislative will itself is evil for prioritizing actions on the basis of sensuality rather than rational duty to respect the moral law. Thus, an evil person is not evil according to what the person does in action but due to the kind of *Gesinnung* the person develops via their meta-maxim.

Similar to Plato's intellectualist account, Kant approaches evil agency with the view that morality is inseparable from its rational core. As such, evil must be to some extent a corruption or defect in human beings' rational faculties. In cases of perverse evil, agents are generally ensnared by the incentives associated with self-love and believe them to be good in some way. Their maxims are, at bottom, hypothetical imperatives that are conditioned by desires (or circumstances) and are not properly moral. The passions are pathologically subversive to the categorical imperative's decision procedure and often work in opposition to coerce agents away from maxims that adhere to the moral law. Hitler, according to the parameters of Kant's

[144] Allison, *Kant's Theory of Freedom*, 122.

[145] This allows Kant's internalist model to rather easily explain instances of akrasia, amorality, and some evil characterizations. Section 3.1 will briefly cover the matter. The most important question, though, is whether Kantian internalism can address purely evil agency as defined throughout this inquiry.

internalism, would be evil in the sense that he based his moral judgments on considerations in direct opposition to the moral law. His judgments were presumably rooted in incentives that were desirable to him rather than towards fulfilling the moral law. From the normative standpoint, one could appeal to the categorical imperative which demands that maxim (i.e. judgment) formulation be regulated by considerations of universalizability, the ends rather than the means of conduct, and the autonomy of all other rational agents. But metaethically, in terms of moral psychology, perversely evil agents have corrupted their principle of maxim-selection before even making a moral judgment or acting on it.

The case of Anakin Skywalker, described earlier in section 1.1 as a peculiar example of perverse evil, can be interestingly construed from the Kantian approach. When Anakin was first submitted as a candidate for the Jedi order, the Jedi council headed by Yoda had reservations about whether or not he could be safely trained. The Jedi analysis offered by Yoda in *The Phantom Menace* (1999) is not that different from the Kantian interpretation of (perverse) evil. The Jedi, according to Yoda, "must have the deepest commitment, the most serious mind" and as such require the proper control (i.e. prioritization) of one's appetites and emotions in relation to one's reason.[146] The life of a Jedi, like that of an ascetic monk, requires strength of character and reverence for one's duty to good principles. Given the extraordinary powers one obtains being trained to use the Force, there are dire consequences when one adopts a principle of maxim-selection that is not conducive to a rational duty for duty's sake. If one's maxims are established on the basis of self-love, then it is not too difficult for maxims to deviate from the categorical imperative due to the moral law's de-prioritization. Kant, like the Jedi, acknowledges the power of sentiment in our daily lives and choices. Having feelings by themselves is not dangerous; rather it is when the passions serve as the moral ground of actions that makes them dangerous. One could interpret the Jedi mindset as one that has the fortitude to repel the incentives of self-

[146] *Star Wars: Episode 1 - The Phantom Menace*, directed by George Lucas (Marin County, CA: Lucas Films, 1999), DVD.

love and choose a life of duty and compassion, whereas the Sith reveling in power and passion opt for self-love.

Given this conceptual framework, Kant does not encounter the problems that seemingly plagued Socrates and other similar internalist accounts in section 2.1 and 2.2. Kant's internalism, instead of pointing to the actions themselves, designates Hitler's will (via the Incorporation thesis) as evil. Genocide, greed, and maliciousness are things that are made possible by adopting a kind of mindset that would find those things motivating in the first place. While Hitler's actions are not indicative of an evil will, for one can have an evil will and have actions legally comply with the moral law,[147] the motivating influence of his desires over and above any properly rational consideration makes it possible for the incentives of self-love to be subjectively chosen and notably evil actions can arise given the right conditions. Pablo Muchnik points out that the evil Gesinnung essentially underlies one's maxims with the subjective determination that "I will what I please."[148] Failure to prioritize the moral law can be explained in ways that have already been addressed previously: perhaps due to complete ignorance (e.g. idiocy or imprudence), overriding desires (e.g. obsession/mania), or some other circumstance (e.g. upbringing or education). The label of self-love, for Kant, is merely a general principal of maxim-selection and can cover a wide variety of phenomena. This may explain Hitler's case and certainty many others, but what about pure evil?

Concerning purely evil agency, Kant—in some ways similar to Anscombe and McNaughton's assessment of Milton's Satan in section 1.2—rejects the conceptualization of evil qua evil motivation. It represents a "corruption of [one's] morally legislative reason" which is

---

[147] Kant, *Religion*, 53-54 [AK 6:47].
Kant's own explanation here is quite clear: "When the firm resolve in complying with one's duty has become a proficiency, it is also called *virtue* in terms of legality, as virtue's *empirical character*...However …[one] should become a human being who is not merely *legally* but *morally* good…who requires no other incentive beyond this presentation of duty itself."
[148] Muchnik, *Kant's Theory of Evil*, 108.

inapplicable to human beings.[149] While perversely evil agents succumb to the dictates of self-love and reverse their order of incentives to reflect this priority, purely evil agents operate on a malignant reason opposed to the moral law in which it is completely renounced. This thoroughly evil will affirms the meta-maxim of self-love but goes much farther than merely "I will what I please". In effect, the purely evil agent immediately rejects the moral law for what it is and then inverts the motivational power of the moral law as the basis to commit evil. In other words, the agent does not choose evil qua some perceived good or desire via "the incentives of self-love" but rather opts to do evil because of the fact that it is contrary to the moral law. The difference here is a crucial one. Unlike perversely evil agents, the purely evil agent is not motivated to prioritize self-love over the moral law because the motivating incentives of self-love have triumphed over the incentives of the moral law. As discussed extensively in previous sections of this thesis, this conceptualization of evil intentions escapes the label of perverse or instrumental evil since the agent in question does not couch their pursuits in relation to some closeted good. Additionally, this agent cannot be rightly called amoral because he/she does care about morality—just not the side of the moral issue one expects.

However, Kant in *Religion* does not seem consider purely evil agency to be distinguishable from ordinary (i.e. perverse) evil. He frames evil for evil's sake within the domain of the diabolical—as belonging to some other otherworldly being.[150] Based on Kant's a priori considerations above, the human being is limited to the extent that one cannot excise the moral law and coherently retain one's rational nature and sense of freedom. Kant's rejection of diabolical or pure evil is fueled by his a priori considerations outlined above in *Religion* and his other works. Though agents may prioritize the incentives of self-love over the moral law, they cannot be so completely dedicated to self-love as to render the moral law empty; it must remain

[149] Kant, *Religion*, 39 [AK 6:39].
[150] Ibid.

as a possible configuration of the human being's Gesinnung.[151] To be human is to have at least some remnant in which the moral law can be motivating. Thus, human beings cannot be purely evil because they cannot eradicate the moral law without consequence (e.g. losing one's own humanity).

As section 2.2 outlined, pure evil agency seems to turn the MIT on its head. The purely evil agent, as defined throughout this inquiry, seems to precisely counteract the idea that an agent's moral judgment of what is good contains sufficient motivation to act on it. Kant, though, uses an a priori framework in which the moral law stands above repudiation and even the most evil of human beings still possesses some natural predisposition to the moral law. The exception to this—the diabolical or purely evil will—cannot be embodied within human beings. To choose evil-qua-evil, the purely evil agent upon recognition of the moral law must immediately renounce the moral law on its own basis. To do so would be a corruption of the very source of reason which the moral law makes possible in the first place. Such an incentive, according to Kant, is rooted in a malignant reason and cannot be exercised by a human will.[152] While human beings have the freedom to choose either a good or evil will, there is never a time where either incentives of one are completely eradicated. The moral law, no matter how much it is de-prioritized, always remains in the picture. Thus Kant's internalist account of evil seems to escape the difficulties that were presented in the previous chapters by preserving the concept of pure evil but regulating it to other forms of agency distinctly non-human.

There is more to Kant's account that needs to be explored. The basic outline of Kant's approach in this section subtly brings together the concerns of the first chapter and the nuanced, theoretical details of the second chapter. But there still remains a lingering sentiment that pure evil—or something that closely resembles it—is realizable for certain agents while also retaining

[151] Ibid.
[152] Ibid.

their humanity. The use of literature in sections 1.1 and 1.2 was instrumental to illustrating the utter flexibility that evil characters can have. Some of these characterizations brought to light another dimension of evil agency that seemed to mirror the opposite of moral sainthood, a kind of principled stance towards being evil. The incentive to destroy that which is good for the sheer reason of reveling in the act for itself does not seem beyond human capacities but in some sense is at the very core of the concept of evil in moral discourse—especially when reflecting on the depths in which Dostoevsky and others portray humanity. One must, contrary to Kant, take this phenomenon as it stands and seek out ways to make sense of it. Is there a way to salvage some instantiation of pure evil within motivation internalism as well as Kant's theory of moral agency that connects with both the imagination and the complex motivational depths of man?

CHAPTER III

Kantian Evil

The previous chapter gave a basic understanding of the motivation internalism and externalism debate. Both metaethical positions explicitly oppose one another on the issue of whether an agent's moral judgments are necessarily motivating (irrespective of the agent's subsequent actions). While the externalist argues that an agent can make judgments and be entirely unmoved by them, the internalist maintains that judgments by their very nature convey a "pro-attitude" that impacts one's actions. Even if one's judgments are undermined by other motivations or circumstances, the internalist can still appeal to the original motivation to form the judgment to begin with that failed to be carried into action. Or perhaps the judgment itself was not genuine in the first place. The bulk of chapter two outlined these explanations and some implications on moral agency. The introduction of purely evil agency complicated this debate even further.

Immanuel Kant's approach, briefly outlined in section 2.3, provides an a priori basis for the MIT and furthermore posits an account of evil that conceptually recognizes evil-qua-evil motivation but at the same time rejects any human exemplification of it. Kant's account presents a remarkably structured understanding of evil that avoids much of the issues that plagued chapter two. But a major concern still remains. Can motivation internalism not just account for the concept of purely evil motivation but also make room for the conceptual possibility of purely evil human beings? Though Kant seems to reject the hypothesis, this chapter will explore whether there is a gap within Kant's internalism for human beings to express purely evil motivations.

Section 3.1 will outline Kant's three grades of evil that arise out of human beings' general propensity towards evil. These different levels of evil relate in many ways to the various instances of agency discussed earlier—amoral, akratic, and perversely evil agency. Section 3.2 will address Kant's view of the diabolical with reference to his three grades of the evil propensity in humans. Also, diabolism will be contrasted with this inquiry's conceptualization of purely evil agency in order to demonstrate critical differences. These critical differences between the diabolical being and the human being with evil-qua-evil motivation will be utilized in section 3.3 in order to see if there is room in Kant's a priori account of evil for evil-qua-evil motivation as a conceptual possibility for human beings. Appealing to the views of Paul Formosa and Irit Samet-Porat, it will be argued that Kant's account of evil does have room for human agency that can pursue evil for itself. Without undermining Kant's rejection of a diabolically evil human being and at the same time giving humanity the potential to realize destructive and principally evil tendencies, this modified view of Kantian evil seems to further strengthen the motivation internalist position explored in chapter two. These considerations may offer this inquiry the insights it seeks.

3.1 The Three Grades of Evil

Immanuel Kant's a priori analysis in *Religion* derives evil from the heterogeneous structure of the human will, a result of the freedom to legislatively prioritize the incentives of self-love over the moral law. Section 2.3 outlined some of the important components of Kant's internalism and the identification of evil as originating in the agent's Gesinnung (disposition). However, there are psychological nuances to the way self-love can manifest in the human creature once it is incorporated as the ultimate principle of maxim-selection. Just as section 1.1 established multifarious parameters to being evil, self-love also admits of degrees or grades. This ranking of evil offers a certain depth to Kant's internalism in which amorality, akrasia, and most perverse evil can be explained without much difficulty. Surveying Kant's three grades of evil will be important for later sections as there may be a gap or opening in his views for enigmatic human beings to take the principle of self-love to new extremes—such as the prospect of evil-qua-evil motivation first introduced in section 1.1.

The three distinct grades of evil described by Kant in *Religion* are frailty, impurity, and wickedness (or corruption) of the human heart.[153] Each grade represents a particular way human beings can de-prioritize the moral law. The first, frailty, quite explicitly covers instances of akrasia. Kant references the lamentations of the Apostle Paul ("Willing I have indeed, but perform the good I cannot!") to highlight the agent suffering from a frail will.[154] As discussed extensively in section 2.1, the akratic agent is not ignorant of what is deemed morally good. Neither is the agent willfully drawn to do what is considered evil. Put in Kantian terms, the objective apprehension of the moral law does not guarantee that an agent will subjectively incorporate the relevant meta-maxim(s) into their will. In other words, acceptance of the moral

[153] Kant, *Religion*, 32 [AK 6:29].
[154] Ibid.

law does not assure an agent's motivation to act in that fashion.[155] The human will, being susceptible to frailty by every kind of condition, can be enticed away from adhering to the moral law. Some of these conditions have already been discussed in section 2.1 (e.g. depression, addiction, rage).

The second grade of evil, impurity, identifies agents with a will that does not adhere to the moral law for *duty's* sake. One performs what is deemed morally good for some other sake—perhaps due to some external influence (e.g. "God wants me to do 'X'") or utility (e.g. "It is in my self-interest or for the greater good that I do X")—whereas the good will "admit[s] the law *alone* into itself as *sufficient* incentive."[156] This grade designates not just some of the perversely evil agents described in section 1.1 but also encompasses most agents at some point or another—even those in moral discourse one would presumably call good. To make moral decisions prudentially based on circumstance or feeling as opposed to principally based on reason and duty is the mark of impurity. As Kant similarly argued in *Groundwork*, an action ceases to have genuine moral worth if the motive to act on it is conditioned by some desire or feeling.[157]

A utilitarian, for instance, may determine (via the Greatest Happiness principle) the appropriate action that corresponds with the moral law based on considerations about the welfare of the agents in the situation and the outcomes that would likely result from said action. The utilitarian decision-procedure, though, is impure because the action is not approached from the unadulterated motives of duty. Like Bernard Williams' Jim and Indians case discussed in section 2.1, the utilitarian determination of what is good can easily give way to circumstances that dictate

[155] In relation to Socratic internalism, which seemed to struggle with the motivational strength of an agent judgment towards good and evil, Kant's view is a step forward in line with many contemporary internalists. This distinction between the Socratic view and others is often designated as strong internalism and weak internalism to prevent judgments from being never overriding by some other factor(s). See Shafur-Landau, "A Defense of Motivational Externalism," 267-268.
[156] Kant, *Religion*, 32 [AK 6:30].
[157] See Kant, *Groundwork*, [AK 4:397-399] for illustration of four kinds of motivational grounds of action—with the fourth being the example of a good will.

violating one's own integrity or those of others—even if the conditions may be extremely unlikely. While the utilitarian may regularly judge and act in accordance with the moral law as legally moral, the intention of incorporating the moral law as a rational duty and priority within one's incentives is neglected. There is more to morality and being a morally good person for Kant than simply doing the right thing. Having the appropriate motives establishes the a priori principle that determines one's basis for maxim-selection. Adopting a view that emphasizes cost-benefit analysis of actions undermines duty to the moral law as a priority in one's incentives.

Perhaps one can go so far as to say that the utilitarian ethic—like most consequentialist ethics—lends agents to develop an evil Gesinnung, whereas an ethic of duty aids toward the development of a good Gesinnung since one's motivating reasons for action would not be far removed from Kant's deontological ethic.[158] Any agent that tries to include an extra incentive within their duty to the moral law subsequently undermines the moral worth of his/her actions. The unadulterated performance of duty for itself demonstrates the unqualified commitment of an agent's good will and preserves the agent's respect for the moral law as independent of other inclinations. In any case, it is not enough that an agent performs a morally good action, but also that the action is done with the proper order of priority within one's incentives for action (via the a priori principle of maxim-selection).

Kant's motivation internalism is informed by his normative theory of ethics. The utilitarian has adopted incentives that have diluted rational respect for the moral law in favor of cultivating reasons dependent on some given condition(s). Utility is not only irrelevant to moral evaluation, according to Kant, but also can readily be a pathological obstacle to fulfilling one's

[158] It is important to note though that there is still much that can go wrong when an agent adopts an ethics of duty. For Kant, the duty must emanate from a rational respect for the moral law. Other forms of duty based on social, political, or cultural conditions are not equitable to one's moral duty (but may be integrated as imperfect duties).

duty for duty's sake.[159] As such, the impurity of the human will can occur when an agent arrives at moral judgments on the basis of hypothetical or conditional reasons rather than categorical or non-conditional reasons. The moral law ceases to be a sufficient incentive for action when agents put their feelings or desires on par with (or supersede) duty. One need not explicitly or consciously prioritize the inclinations of self-love to forsake one's rational duty to the moral law.

There is a crucial difference, in Kant's estimation, between the person who performs good actions because they feel right and the person who performs good actions because they conform with a rational duty to the moral law. Even if such agents do not explicitly perform evil or immoral actions, they have established a motivational foundation that ultimately lacks proper moral force. A human being of good morals can possibly live a life to the letter of the moral law, but to be a morally good human being necessitates also adhering to the "spirit" of the moral law as "the sole and supreme incentive."[160] Agents can have their judgments (and consequently their actions) be compliant with what the moral law dictates, but such judgments require a will that has incorporated the proper principle of maxim-selection in order to be considered morally good.

Moral judgments cannot find permanence within incentivized appeals to desires or prudential circumstances—what Kant categorizes as hypothetical imperatives. The statement "If you want to avoid going to jail, then you should not (among other things) steal other people's possessions" is persuasive only if an agent affirms the logical relation between the antecedent and the consequent of this statement and finds the desire (motivationally) compelling to warrant compliance. There are likely agents that do not care whether they are caught or are not threatened by the prospect of discovery. Moral considerations based on hypothetical imperatives could easily evaporate if prudence fails to motivate such an agent. Kant's second grade of evil here seems to signify the tendency of the human will to value and misattribute morality to prudential

[159] See Kant's example of the benevolent philanthropist in *Groundwork* [AK 4:397-398].
[160] Kant, *Religion*, 33 [AK 6:30-31].

reasoning. Thus, while impurity is a lower grade than wickedness, there is a sense in which impurity could naturally lead over time to further corruption.

The third and final grade of evil is wickedness when an agent's will directly incorporates the incentives of self-love—hence establishing an evil Gesinnung. Such a disposition orients the agent's judgments to openly act contrary to the moral law. For Kant it represents the zenith of evil in the human creature because the agent has established as the basis for determining action incentives that are utterly corrosive to individual autonomy as a rational being. Whereas the second grade of evil, impurity, designates an agent that may briefly flirt with the incentives of self-love as a means to objectively fulfill the moral law, the corrupted will makes its home in self-love as the preferred order of priority for deciding how judgments are formulated. The moral law takes a back seat as other inclinations are given consideration. Pablo Muchnik makes similar remarks in the following passage:

> "Although in frailty and impurity the moral incentive does not receive authority in the motivational structure, the agent makes at least a lukewarm attempt to acknowledge the outweighing character of moral reasons. In perversity, all pretenses fall off: the agent unabashedly places herself above the law. Instead of hiding and justifying the inversion of the order of priorities, like in other propensities, this type of agent willfully embraces it."[161]

While this might sound very similar to the characterization of pure evil described in previous sections, for reasons that will be explored later Kant dismisses the possibility of human beings embracing self-love for itself. He refers to such evil as diabolism and quickly distances it from wickedness or any other grade of evil. The diabolical agent extirpates the moral law while the wicked merely dethrones the moral law from the order of priority in favor of other, contrary, inclinations. Thus, human beings are conceptually limited a priori on the extent self-love is incorporated into their legislative will. To understand the limitations of the human evil in this

[161] Muchnik, *Kant's Theory of Evil*, 161.

capacity, one should consider of how all three grades of evil establish limitations on the human will to choose evil.

Regardless of whether the moral law is even consulted at all, it nonetheless remains in the background within all three grade of evil—even the wicked agent can only de-prioritize the moral law but never eradicate it as a potential choice. This critical insight demonstrates an important aspect of Kant's version of motivation internalism, which underlies his rejection of diabolism for human beings. However much a human being may rebuke or revile the moral law, there must still be some "germ" of original good that one is receptive to (what Kant designates as a predisposition to humanity).[162] That is, the possibility of the moral law as a motivating incentive for the basis of one's principle of maxim-selection (or meta-maxim) must be present for any agent to have the power of choice. Otherwise, allowing the incentives of self-love too much persuasive force over the moral law (insofar as one can renounce it entirely like the diabolical agent can) would undermine Kant's sensitive notion of freedom. Lawrence Pasternack makes the same observation below:

> "Kant needs his account of evil to find a middle course between diabolism and unintentional immorality. He is neither willing to accept the possibility that an agent can directly reject the moral law, nor is he willing to mitigate freedom through heightening the power of sensible inclination."[163]

The latter is particularly relevant to this section as it relates to the amplification of each successive grade of evil towards, but never quite reaching, the extreme of diabolism. How does Kant limit and curve the power of sensible inclination in his three grades so to prevent any agent from transcending to the diabolical? Kant's answer, as understood among most Kantian scholars, is "self-deception."[164]

---

[162] Kant, *Religion*, 30 [AK 6:28].
[163] Lawrence Pasternack, "Can Self-Deception Explain Akrasia in Kant's Theory of Moral Agency?", *Southwest Philosophy Review* 15, no 1 (2000): 93.
[164] See in particular Allison, *Kant's Theory of Freedom*, 159-161.

The first two grades are marked with self-deception based upon the lack of intent that both have towards prioritizing the moral law. The frail agent is deceived about the motivational strength of the moral law and resigned to give in to the incentives of self-love. Muchnik describes the process of self-deception for the frail agent as the rationale that "Since I incorporated the good (the law) into my maxim, any deviation is not really due to my evil heart, but to the weakness of will, of which I am not entirely responsible, since frailty is part of human nature."[165] The frail agent is neither fully committed to self-love nor the moral law, but nonetheless grants primacy to the incentives of self-love due to a perceived lack of strength. As such, the agent's objective awareness of the moral law falls short of being subjectively motivating when stronger inclinations come into the picture.

Recall earlier that section 2.1 described instances of depression and addiction as typical cases of akrasia or weakness of will. The psychological well-being of an agent as well as the capacity for moral deliberation, which would be otherwise operating normally, can be severely strained given these harrowing conditions. While the person suffering from depression or addiction may insist or make it seem as if the choice is outside of their control, Kant would maintain that the dominance of inclination "can only be the result of the will's 'taking' the sensible incentive as motivating …in terms of the comparative weakness of the moral incentive."[166] Such agents can be indicted as nonetheless making a choice to submit to the perceived stronger inclination when, for Kant, there is sufficient incentive already to adopt the moral law. The frail agent hides making a choice in favor of self-love as one's meta-maxim in this instance; and under the guise of a weak heart, the agent utilizes self-deception in thinking that he/she lacks inner strength to repel seemingly overwhelming conditions.

"…[I]t is also necessary to assume that self-deception, which Kant only mentions in connection with the third degree, or intentional guilt, is operative from the beginning and, indeed, is an essential ingredient in the propensity to evil."

[165] Muchnik, *Kant's Theory of Evil*, 157.

[166] Ibid.

The impure agent, on the other hand, is self-deceived by means of believing one is acting with the moral law in mind when it is actually an inclination inspired by the principle of self-love. An agent may act in accordance with the moral law, but it is done for the wrong reasons or motives. The agent's good actions are incidental rather than being representative of motive of a good will. The person who refuses to steal because it displeases others acts on a motive that is dependent on wanting to fulfill that desire (of not displeasing others). The person who refuses to steal because it is a violation of duty places morality on an unconditional ground over and above sensuous inclinations. Whereas the frail agent is objectively aware of this duty to the moral law and fails to act on it due to some perceived (i.e. imagined) weakness of will, the impure agent's self-deception is simply one of error and ill-conceived ignorance. Though the impure agent may do what is good without hesitation, he/she does not have an "attitude within the law of duty"[167] which signifies a properly moral understanding.

One cannot help but think that John Stuart Mill's utilitarianism, a theory which Kant would presumably oppose, is unintentionally guilty[168] of diluting the moral law to incentives that are contrary to designating genuine moral worth on one's actions. The fact that the Kantian good will and Mill's "impartial and benevolent spectator"[169] could produce similar judgments in moral situations is merely incidental. Utilitarian moral reasoning, as Bernard Williams argued earlier, is relative to the situation such that lying, stealing, or perhaps even murder may be obligatory in order to satisfy the greatest good for the greatest number. This is again because of the distinction between hypothetical and categorical imperatives highlighted earlier in Kant's *Groundwork* and

[167] Kant, *Religion*, 41 [AK 6:38].
[168] Ibid, 41-42 [AK 6:38].
Kant regards both frailty and impurity as propensities that unintentionally or inadvertently guilty of giving credence to the incentives of self-love. However, the third grade of evil (wickedness) is intentionally or deliberately guilty in its adoption of self-love as the principle of maxim-selection—but not so much that evil itself becomes more preferable than the moral law.
[169] John Stuart Mill, *Utilitarianism*, ed. Oskar Piest (Indianapolis: Bobbs-Merrill, 1957), 22.
"As between his own happiness and that of others, utilitarianism requires him to be as strictly impartial as a disinterested and benevolent spectator."

other writings. The utilitarian, as an impure agent, has deceived himself into thinking that his deliberations and motives are in accordance with moral duty when they are in fact superficially conditional and can readily evaporate (or be overridden by other circumstances) in an instant. Kant's identification of moral goodness as a principled grounding gets to the core of his three propositions of morality in the *Groundwork*.[170]

The wicked agent takes impurity to its next level and, for Kant, the final manifestation of the evil Gesinnung attainable for humans. Wickedness encompasses a fully matured self-deception that outright neglects the moral law and directly embraces the reversal of incentives that prioritizes the principle of self-love as one's meta-maxim. Kant quite clearly labels wickedness as a "corruption and perversity of the human heart."[171] As such, the agent has systematically deceived oneself in the worst way possible. Muchnik concisely describes this in the following passage:

> "An agent with a depraved heart does not simply transform morality into a system of conditional imperatives, but perverts moral judgment at its root. In depravity, deliberation becomes oblivious of morally salient features that accompany actions, and such insensitivity opens the possibility of maximum wrongdoing…Although [the agent with a depraved heart] is aware of what morality requires, she grants herself 'moral holidays' and callously uses everyone else as a tool to her goals, justifying her conduct in terms of a preserve [*sic*] conception of the good."[172]

The systematic self-deception presented here is more corrosive to the agent than in the previous two grades. Instead of an agent (impurely) mitigating the dictates of the moral law through sensuous inclinations, the wicked agent adopts the incentives of self-love at face value and deceives oneself in thinking that one's conduct is good. As such, the wicked agent is at the same footing as the perversely evil agent with a twisted understanding of good whereby the term

[170] These propositions can be adequately summed up as the following: (1) A good will acts from duty, not from inclination [AK 4:398], (2) "An action from duty has its moral worth not in the purpose to be attained by it but in the maxim in accordance with which it is decided upon" [AK 4:399], and (3) "duty is the necessity of an action from respect for law" [AK 4:400].
[171] Muchnik, *Kant's Theory of Evil*, 33.
[172] Ibid., 160-161.

"good" is disassociated with any real or imagined sense of duty. The impure agent still respects his/her duty to fulfill the moral law, but fails to grasp its proper form (e.g. the utilitarian). The wicked agent, though, forsakes duty entirely and acts on (selfish) impulse.

Kant does not want to designate wickedness as malice.[173] In part, this hastily associates acts of cruelty and viciousness with an agent's Gesinnung (i.e. disposition or attitude). Like impurity, wickedness can be compatible with lawful moral action but "no matter how virtuous someone may be, whatever good he can do is yet merely duty."[174] A change of heart is required within one's Gesinnung to be morally good. Kant describes malice in the strict meaning of the word as a taking evil-qua-evil into one's maxims.[175] If so, the agent would not be (self-) deceived about choosing evil as good and instead relishes what is evil for itself. This would make the agent diabolically evil rather than wicked.

While this grade of evil is a step above impurity, wickedness is not choosing evil for its own sake either. The moral law is (deceptively) ignored or dismissed in place of something else, but never eliminated as a possible choice. Self-deception acts as a limitation on all three grades of evil in order to preserve the freedom that humans have between animal impulse and rational determinism. If a wicked agent ceased to be self-deceived and incorporated evil as evil into their maxims, then the agent would become diabolical and be rendered non-human—a conceptual absurdity. Such maliciousness in one's Gesinnung cannot be maintained by any human capacity to reason without falling into absurdity and contradiction.

The basic reasons for Kant's rejection of the possibility of diabolical evil for human beings have already been introduced in section 2.3. It was remarked then that Kant's internalism

[173] Kant, *Religion*, 41 [AK 6:37-38].
[174] Ibid, 33 [AK 6:30-31].
[175] Ibid, 41 [AK 6:37-38].
"…if one takes this word in the strict meaning, namely as an attitude (subjective principle of maxims) of admitting, as an incentive, evil as evil into one's maxims…"

seems to side with Anscombe and McNaughton's views in section 1.2 when discussing the extent of Satan's motivations in Milton's *Paradise Lost*. This section's outline of Kant's three grades of evil seems to further support that, based on the particular degree (i.e. grade) by which evil takes root in an agent's legislative will, any human case of evil-qua-evil motivation must be misperceived as some perverse conception of good or identified as some other end (e.g. evil-qua-power, evil-qua-desire, etc). To say that one can choose evil for itself would be a misnomer since, for Kant, the moral law is necessary for structural autonomy in the first place. Human beings simply cannot be free, or rational, and renounce the moral law without consequence. Given these further details in this section, a comparison needs to be made between this inquiry's conceptualization of pure evil and Kant's account of diabolical evil in *Religion*. Are both notions equivalent? Is it possible that evil-qua-evil motivation can be distinguished from Kant's diabolical evil? Furthermore, is there a way to fit this notion within Kant's internalist account without dissolving his a priori premises that regards diabolism as incompatible with human beings?

3.2 Diabolical Evil

Kant's diabolism in *Religion* draws limitations on human evil. His rejection of a diabolically evil human being is consistent with his internalist model and a priori framework that underwrites it. Human beings, as agents with inclinations that can conform to either reason or the passions, have a will that affords them the choice to set up an order of priority in their judgments and subsequent actions. But, for Kant, there are boundaries that cannot be trespassed so long as one is human. Simply put, the evil that any human being incorporates as the basis of their maxims does not purposefully desire to do evil but rather is under a spell of self-deception. The three grades of evil described in the previous section indicated multifarious ways human beings can have the principle of self-love ground their evil Gesinnung (disposition). But being evil for evil's sake (i.e. evil-qua-evil motivation) is not represented in any of those grades. Perhaps it has no possible human derivation here and is entirely within the domain of the diabolical? This stance needs to be explored and contrasted with the conceptualization of pure evil provided in section 1.1. Ultimately, this section will not only distinguish between Kant's diabolism and pure evil but also will suggest that evil-qua-evil motivation is not necessarily tied to the diabolical alone.

To briefly summarize Kant's rejection of a diabolical agent, he argues that thinking about a human being choosing evil for itself is "tantamount to thinking a cause operating without any laws."[176] If humans have freedom of choice by which some actions are imputed over others, then such agents must have another option that ensure judgments are freely made and not causally determined one-sidedly. In other words, neither self-love nor the moral law can overwhelmingly fix the individual's choice on the matter. There must be more than one incentive to incorporate as a principle of maxim-selection. Otherwise, the idea of a free choice is rendered absurd because the deck is stacked to where only one option determines an agent's motivational allegiance. As a

[176] Kant, *Religion*, 39 [AK 6:35-36].

result, there would be nothing to demarcate and impute free choice within an agent's judgment (and actions).

By exclusively incorporating self-love and expunging the moral law as a possible incentive in itself, the diabolical will represents a one-sided commitment to the principle of self-love. The imputation of this agent's actions as a product of free choice becomes null and void. As such, freedom is restrained (i.e. limited) because actions necessitate some "mark of agency or authorship"[177] without which agents cannot inscribe their judgments as freely chosen.

Muchnik, earlier in section 2.3, explained this in terms of the "human in-between" whereby free choice is found between the animalistic impulses of passion and pure reason underlying the moral law. Human volition is "characterized by its structural heterogeneity…were the human will directly determined by sensuous impulses, it would be animal; but were reason always sufficient to move it, it would become holy."[178] Diabolical evil upsets this delicate balance and undermines the heterogeneous structure of the human will. Agents would be one-sided or univocal in their judgments. There is not a capacity to do otherwise or the possibility for a change of heart. Like the animal and divine wills, the diabolical will is homogenously structured. The thought process by which evil is "chosen" for itself portrays a mechanical drive absent of Kant's a priori parameters on freedom. As a result, the conditions of diabolical evil stand starkly opposed to every notion of what it means to be human: an autonomous, rational, and deliberative being.

Kant's "Reciprocity thesis"[179] also attaches the same concerns to the rationality of agency. Human beings, for Kant, are dependent upon the existence of the moral law because the very structure of reason would be incomprehensible without it. The moral law establishes a

[177] Muchnik, *Kant's Theory of Evil*, 8.
[178] Ibid., 9.
[179] See section 2.3 for more of an introduction to this underlying Kantian principle.

ground by which the agent's will can autonomously subject itself to laws. Without the moral law at all there cannot be any free action, but merely a lawless anarchy. The lack of structure or order in an agent dissolves any intelligible causality that can be imputable to action. Hence, the diabolical will is contradictory due to the fact that it seeks to renounce the moral law using the power of autonomy and reason (malignantly) against its own causal origin. Muchnik explains diabolical volition in terms of a "self-defeating motivational structure" that "deprives itself of reasons for action."[180] This is why self-deception is—and must be—a key feature of each grade of the evil propensity because it seems to be the only description that can coherently maintain the "in-between" of human freedom.

Diabolical evil is immune to self-deception since the will legislatively expels the motivational efficacy of the moral law and affirms the principle of self-love for itself. As previously explored in sections 1.1 and 1.2, such an agent would be doing evil as evil (i.e. evil-qua-evil) rather than a perverse view where evil (i.e. self-love) is perceived as good. Whereas the three grades of evil can be attributed to some form of self-deception, the diabolical agent possesses a supra-human (as opposed to superhuman[181]) nature that can both deliberately and willfully do evil without being deceived. This is quite different than wickedness which does evil under the prospect that it is actually good in the same way Milton's Satan proclamation "Evil, be thou my Good" was considered a sophisticated perverse evil.

[180] Muchnik, *Kant's Theory of Evil*, 116.

[181] The difference in the way the two terms are being used here is merely a semantic one. The point is that someone who is "superhuman" can still be said to be human, only with additional powers that have expanded, transformed, and amplified certain abilities. To be "supra-human" means, to borrow from Nietzsche's vernacular, that someone has overcome or transcended their own conditions. That is, to become something else not entirely human. The fact that Spiderman can climb walls and produce silky spider webs hardly affects his own status as a human being because he still retains the basic qualities of being human. This would be, for Kant, the internalized heterogeneous structure of the human free will. Someone like Dr. Manhattan from *The Watchmen* (2009), however, is supra-human because his transformation resulted in becoming an altogether different entity with little connection to his former human self and the limitations exclusive to human nature.

Also mentioned in section 1.2, Kant's rejection of a diabolically evil human being seems to be indicative of rejecting evil-qua-evil motivation as a possible human pursuit. An agent cannot be human and affirm evil for itself. Or at the very least perhaps, one cannot retain any semblance of humanity with such a view.[182] If a philosopher or some creative writer presents a case to the contrary, then they simply do not understand the a priori concept of pure reason and the ways in which human beings are respectively limited by the Incorporation and Reciprocity theses. No matter how nuanced empirical evidence suggests the human potential towards destructive and devilish behavior, Kant's a priori premises indicate that the legislative capacities of the human will must be deceived in some manner to affirm maxims that are contrary to the moral law. As such, one cannot be both diabolically evil and human at the same time in the same respect. Either the agent in question seems to be a diabolically evil human being but is in fact an otherworldly demon (i.e. non-human) or the agent is human and deviously perverse or wicked.

Due to the above considerations, one might insist that purely evil agency must also be included as conceptually equivalent to Kant's diabolism. Does this require Kantians, like other internalists, to relegate pure evil within the domain of perversity (at least when it comes to human beings)? Section 3.3 will have more to say on this question. But first is there a difference between purely evil agency and Kant's diabolical agent? Is diabolism a requisite for agents to have evil-qua-evil motivation?

In section 1.1, pure evil was defined as doing what is considered evil based primarily or exclusively on a principled notion of being evil—mirroring opposite of the saying "Be good for goodness sake". In other words, a purely evil agent is someone who does evil because of the fact

[182] It is intriguing to speculate whether Kant would agree that diabolism is inherent to a being or whether it can be the result of a long-term erosion of the moral law into nothingness. There does seem to be a sense in which the moral law's presence as an incentive can diminish over time if attention is not paid to it. Of course, it needs to be mentioned that Kant would likely refuse to consider the moral law being erasable at all, especially for human beings. Not forgetting the religious elements in Kant's work being discussed here, his argument for a human predisposition to good is similar in tone to the Apostle Paul on the matter—that the moral is "written in the hearts" of all (rational beings).

that it is evil. Unlike Kant's wicked agent, the purely evil agent does not make evil out to be good in some twisted or perverse context. Just as someone may view that there is an inherent value to being a morally good person, the purely evil agent thinks there is an inherent value to being a morally evil person. While the latter view may sound strange and unintuitive, the former view is a popular way for many laypersons and even some philosophers to think about moral questions such as "Why be good?" One example is Colin McGinn. In his book *Moral Literacy*, his explanation for why one should be a morally good person begins with the following:

> "What reason is there for being a good person? The answer is, there is no reason—or no reason that cuts deeper, or goes further, than the tautology "because goodness is good". The reason you should be virtuous and not vicious is just that virtue is virtue and vice is vice…Moral justification, like all justification, comes to an end somewhere. At some point we have to simply repeat ourselves."[183]

If McGinn's reasoning holds some significance for those in moral discourse, then perhaps the tautology "because evilness is evil" can conceivably constitute as justification for being an evil person. The purely evil agent's motives seem no different but are morally inverted to vice and being a morally evil person for its own sake instead. There may be instrumental goods to being evil (e.g. pleasure, power, etc) just as there may presumably be expected perks for being good, but both the moral saint and purely evil agent prioritize their own principled grounds regardless of the beneficial or harmful consequences. The capacity of a moral saint's pure motive to do good for itself seems to solicit its reversal, an opposing capacity to do the contrary.

Section 1.1 addressed the difficulty in identifying a concrete example of a purely evil human being. While some possible candidates were suggested, the difficulty still persists. Does this mean that the purely evil human being is as illusory (and unsubstantiated) as a diabolically evil human being? But one could perhaps suggest the same about Kant's good will in the

[183] Colin McGinn, *Moral Literacy or How to do the Right Thing*, (Indianapolis/Cambridge: Hackett Publishing, 1992), 95. He later on develops an argument that associates moral goodness with inner beauty and well-being, but his initial remarks here seem to come across as the underlying (final) reason for being a morally good person.

Groundwork. That is, there seems to be a similar difficulty in clearly picking out a person of good will that represents the moral law as its sole determining ground. The answer cannot be concluded by any amount of observation because one is dealing with the confines of a person's private thoughts and motivations.

Moral discourse often highlights candidates of moral sainthood like Gandhi, Martin Luther King Jr, and Mother Teresa. But regardless of the weight empirical evidence one may have, there simply is no way to reliably assess the moral status of an agent from the outside looking in. For Kant, it would be a mistake to conflate a good or evil will with the appearance of doing morally praiseworthy or blameworthy actions. The issue of whether someone in their actions has prioritized the incentive to be good for goodness sake above all other interests necessarily admits of imprecision because people's actions hardly tell the full story of their motivations. Literature can perhaps offer some avenues to identifying patterns in agent thought, but there is no way to test one's motivations without also losing the purity of the motive in the midst of analyzing it within action. One cannot, a posteriori, approximate a good or evil Gesinnung—only a priori as a pure motive of agency.

Kant would probably agree that many people have and are capable of conforming to the moral law for duty's sake. But, even if it were the case that such moral perfection were possible (to both do and have first-hand knowledge of), one should be skeptical of whether a human life can constantly maintain that high moral standing. As the first two grades of evil in Kant's account seemed to indicate, frailty and impurity are conditions of the structural heterogeneity of the human will such that every individual constantly struggles with this heterogeneity in their moral deliberations. Even moral saints, by virtue of being human, are susceptible to mistakes and imperfection. But that does not mean one should dismiss or reject moral sainthood wholesale. Should not pure evil be afforded at least the same consideration?

At the very least then the purely evil and diabolical agents respectively share the same definitional parameters of doing evil for itself. But diabolism has one critical feature that separates it from purely evil agency. A diabolical agent, by Kantian standards, is not a human being but rather some being that transcends reason and human volition. Throughout this inquiry, purely evil agency has been defended as applicable to humanity. While this may be merely a matter of semantics, pure evil does not necessarily have to be associated with devilishness or diabolical beings. Thus, evil-qua-evil motivation may not be exclusively part of a diabolical will and human beings can partake in it without being considered diabolical.

At this juncture, there seems to be a major point of difference between Kant's diabolism and purely evil agency. Just as purely good and purely evil intentions reflect each other in terms of methodology, one can recognize the moral saint and its evil counterpart pertaining to opposite ends of the moral spectrum as exemplars. Peter Brian Barry calls this the "mirror thesis," the notion that moral saints and moral monsters mirror each other "insofar as the characters of each are marked by similar aretaic properties…each is an instance of the highest degree of something."[184] For Kant, though, human beings cannot mirror the evil counterpart of a good will to the extreme of diabolism. But Kant's rejection here need not include all instantiations of evil-qua-evil motivation. Diabolism requires more than a pure principled stance to do what is evil. It signifies lawlessness and the lack of a causal structure by which a will can rationally impute free choice in one's actions. As such, the one-sidedness of the diabolical will does not complement section 1.1's conceptualization of pure evil and one would be mistaken to suggest that pure evil is fully encompassed within diabolism.

Whereas the diabolical agent is purely devoid of the moral law's irresistibility and the will is monopolized towards the principle of self-love, the purely evil agent inverts the moral

[184] Peter Brian Barry, "Moral Saints, Moral Monsters, and the Mirror Thesis," *American Philosophical Quarterly* 46, no.2, (2009): 163.

law's motivational value as incentive to do evil rather than renouncing it entirely. In other words, the purely evil agent takes the irresistibility of the moral law as motivation to do its contrary—self-love. Far from extirpating the moral law, the purely evil agent has an immediate objective apprehension of the good and subjectively fixes it as an incentive. But whereas the wicked agent operates on the belief that self-love is good, the purely evil agent approves of the moral law as good for the expressed purpose of prioritizing evil above it. Paul Formosa makes similar remarks below about the possibility of divorcing evil-qua-evil motivation as exclusively diabolical within Kant's account of evil:

> "Kant (or the Kantian) *can* make sense of the possibility that a person might choose evil *qua* evil…The affront to our self-conceited conception of ourselves, the attack on our pride and the swift rebuke to our arrogance dished out by the consciousness of the moral law can (perhaps) lead to a 'rebellious attitude' of resentment and hatred toward the law itself. To act directly upon this hubristic hatred of the law is to choose evil *qua* evil."[185]

The parallel between the purely evil human being and the saintly human being maintains Kant's diabolism but at the same time retains the possibility that human beings can do evil for itself. Just as Kant's three grades of evil outlined in section 3.1 show the progression by which an agent's will can develop an evil Gesinnung, this grading could also conceivably be inverted for evil.

It is not such a stretch of the imagination to conceive of an evil-doer undergoing the same process of Kant's grading but in terms of evil to progressively good in attitude (i.e. Gesinnung) due to frailty, impurity, and corruption. Evil agents can presumably experience moments of weakness, fail to do what is evil due to some other conflicting incentive, and perhaps even experience remorse or guilt over their failure to do what is evil as much as any good agent in the contrary situation. In the midst of doing a principally evil act, the agent can be susceptible to any number of emotions or circumstances that could be seen as overriding. For example, the agent could immediately suffer from depression and become listless to all his/her activities. Perhaps

[185] Paul Formosa, "Kant on the Limited of Human Evil," *Journal of Philosophical Research* 34, (2009): 197.

the agent gets attached emotionally to others such that those relationships impede any evil motivations. This can seemingly lead to other conditions that can gradually convert the evil agent to have a change of heart and have a good Gesinnung instead.

In terms of impurity, evil agents can perform a legally evil action but not according to the "proper" order of priority for the evil Gesinnung. The purity of the evil agent's motive to do the action out of duty to self-love is "corrupted" by other incentives such as pleasure, power, glory, etc. As such, the agent is motivationally disposed towards the instrumental gain or consequences of the evil act instead of the principle of the action itself. The agent no longer does evil as evil but rather tells himself that the action is justifiably good in some evaluative sense. Lastly, an evil agent can succumb to the next grade where "corruption" by the moral law's incentives lead to a change of heart in one's Gesinnung. Self-love no longer is prioritized as the agent's meta-maxim and the moral law takes hold and puts incentives into their proper order of priority.

This is obviously manipulating Kant's grading of evil by suggesting that the moral law for a thoroughly evil agent can be inverted to a completely opposite "evil law" such that a human being can perceive a moral duty is to do what is evil (i.e. contrary to the categorical imperative). Kant himself would likely object to treating evil on the same level as the moral law. For evil to be done as evil requires an utterly malignant reason beyond human capacity. But examining the similarities between the pure motives of Kant's good will and the possible pure motives of an evil will may offer a critical distinction between Kant's diabolism and section 1.1's account of pure evil.

Naturally, one can criticize the mirror thesis and the motivational inversion of the moral law in multiple ways. Some things are being taken for granted and can be questioned. It is not quite evident that to think about moral sainthood requires the antithesis of its opposite extreme—the (im)moral monster. Nonetheless, significant portions of the next section will base its

reasoning on an implicit acceptance of this premise. Before moving on to the next topic, some criticisms will be introduced to highlight the difficulty of this purely conceptual reasoning.

In Plato's dialogue *Phaedo*, Socrates makes a number of arguments that depend on this very notion of opposite values or concepts. Generally called the argument from opposites, Plato's Socrates suggests that "all things which come to be…[come] from their opposites if they have such, as the beautiful is the opposite of the ugly and the just of the unjust, and a thousand other things of the kind."[186] With Cebes as the interlocutor, Socrates produces the following reductio ad absurdam argument: if being dead is the opposite of being alive and vice versa, then neither cannot admit of the other without contradiction. The contradiction is that if the living does not come into being from an opposing status (i.e. death), then there is a lack of balance between the two states of becoming such that everything would cease to be living at some point; but for Socrates this is simply not true.[187]

Additionally, how can one make a distinction in the first place without some comparison? That is, how can one have knowledge of being awake without also being able to distinguish its opposite state? The argument from opposites seems to operate on intuitive grounds. Waking up necessitates a corresponding opposite (i.e. sleeping) just as any concept needs a boundary that delineates it from some other notion. The person who has been asleep their whole life would know nothing of its opposite; nor would the person be able to tell the difference until the contrast presents itself. If Socrates' reductio holds that a concept such as being awake cannot be rendered sensible without some contrasting state (i.e. being asleep), then Socrates has seemingly affirmed his argument from opposites. One can presumably subject the same reasoning for the opposites light/dark, beautiful/ugly, and as the mirror thesis suggests the moral saint and its direct reverse.

[186] Plato, *Phaedo*, [70e-71].
[187] See *Phaedo* [70e-72e] for the complete argument that Socrates establishes with the assistance of Cebes.

Cebes and the other interlocutors in the dialogue seemed to be easily convinced, but there are some avenues for criticism.

Friedrich Nietzsche, in *Beyond Good and Evil*, strongly criticized taking this reasoning seriously as it represented a "faith (in opposite values)" and highlighted the "typical prejudgment and prejudice which give away the metaphysicians of all ages."[188] Nietzsche warns his readers to be skeptical of what seems like a natural thought process. There is a danger in falling prey to language and the way it can manipulate one's thinking.

> "How could anything originate out of its opposite? for example, truth out of error? or the will to truth out of the will to deception?... For one may doubt, first, whether there are any opposites at all, and secondly whether these popular valuations and opposite values on which the metaphysicians put their seal, are not perhaps merely foreground estimates, only provisional perspectives, perhaps even from some nook, perhaps from below, frog perspectives, as it was, to borrow an expression painters use."[189]

There does seem to be something suspicious about Socrates' conceptual rationale and perhaps this relates to the mirror thesis as well. The questionable assumption here is whether concepts as words can take on a meaning without some separation or comparison—as the argument from opposites seems to suggest.

The mirror thesis at first glance seems to greatly resemble Socrates' argument from opposites in the way abstract relations are dependent upon each other for coherence. Everyone has likely heard at least on one occasion the common trope "one cannot have good without evil" in moral discourse. Evil as a privation of the good continues to be a topic that occupies philosophers today as it did during Plato's time and the Scholastics during the Middle Ages. The predicament of evil's existence and intelligibility also extends into philosophy of religion, epistemology, and metaphysics. How can one have a notion of goodness if it is not being contrasted with some opposite state or privation? This difficulty must be overlooked for now.

---

[188] Friedrich Nietzsche, *Beyond Good and Evil*. Trans. Walter Kaufman (New York: Vintage Books, 1989), 10 [sec.2].
[189] Ibid., 9-10.

But Peter Brian Barry's investigation of the mirror thesis in his own article explores many of its nuances that are largely off-topic (but still very intriguing) to this inquiry. He does make one critical comment that is relevant to the Nietzschean criticism above.

> "To be sure, the mirror thesis is hardly adequate as an account of moral sainthood; roughly, it characterizes the structure of an evil person's character but not its content. A full-blown account of evil personhood would also say something about what particular vices, if any, from which an evil person must suffer. Further, the mirror thesis does not explicitly suggest any particular relation between evil personhood and evil action, whatever relation that should be. But any account of evil personhood must start somewhere, and the mirror thesis represents a plausible place to start."[190]

While a "full-blown account of (evil) personhood" is not the primary aim of this thesis, exploring the motivational underpinnings of evil agency could yield content for normative ethical theorizing. If evil acts can be (sincerely) chosen for themselves, then one could ascertain the vices that underlie the agent's will to do what is evil. Like the case of Satan's pride in *Paradise Lost*, vices can be expressed in a pure form untainted by negativity or depravity. Hence, the idea of a purely evil agent is the idea of someone who principally stands for the action in itself. In other words, the action is itself the motive.[191] However, human beings typically perform evil acts with motives that are corrupted by circumstances and oft-misplaced desires—as with various anti-hero and anti-villain archetypes.

This inquiry though must return to the task at hand, to understand evil within the Kantian internalist model of moral motivation. For the time being, without dismissing or ignoring the above criticisms, the mirror thesis will be conditionally utilized for the purpose of advancing to the next topic.

---

[190] Barry, "Moral Saint, Moral Monsters, and the Mirror Thesis," 173.

[191] Consider as an example the Joker from Christopher Nolan's *The Dark Knight* (2008). If one believes his claims, the Joker refers to himself as "an agent of chaos" that does things for the purposes of upsetting the established order. Whereas many villains are motivated by aspirations of wealth, fame, and/or power, the Joker spurns all these instrumental benefits. One scene in particular shows the Joker burning an extraordinary amount of money to the dismay of several mobsters. Of course, one can discount the Joker's sincerity and whether he has ulterior motives that can amount to perverse rather than purely evil motives. But there is at least a distinction here between motives that seem to correlate with this thesis' division of evil.

Due in part to the mirror thesis, there seems to be a possible gap between the diabolical and purely evil. This can have substantial implications for Kant's account of evil as well as the motivation internalism/externalism debate in general depending on how one interprets the limitations of human evil. The next section will explore two possible ways to incorporate evil-qua-evil motivation into Kant's thought. One interpretation, suggestive of the approach above, will suggest that pure evil does not necessarily preclude the moral law. That is, it can be viewed as a pernicious and exceedingly rare grade of evil either as an unconventional form of wickedness or as an entirely new grade of the evil beyond wickedness but not to the level of diabolism. The other interpretation will suggest that Kant's a priori rejection of diabolism is at the root of the problem should be abandoned.

3.3 The Purely Evil Human Being

Kant's a priori account of evil fits well with the concern at the very beginning of this thesis. One should not readily accept the physical appearance of evil-doing and look beyond the images of evildoers in dark alleyways with sinister countenances. For Kant, one must utilize a priori reasoning and delve into how the concepts of reason, autonomy, and morality relate to human beings. The motivational depths of an agent's thoughts and judgments here play a very pivotal role in understanding what it means to be evil. There are constantly principles being prioritized or de-prioritized behind every judgment and action that one observes in moral conduct. Previous sections already outlined much of Kant's account in this regard, but Kant's rejection of the diabolically evil will puts his internalist position at a possible disadvantage comparable to other internalist views in chapter two.

Section 3.1 explained Kant's three grades or levels of evil as a propensity in the will that progressively (absent a sudden change of heart) settles into an evil Gesinnung. But there is a stopping point for human beings. One cannot be evil for its own sake, purely embracing the principle of self-love as the basis for one's moral judgments. If evil cannot be chosen for its own sake without contradiction, then motivation externalists can reiterate its conceptual coherency and their own flexibility on the matter. Such examples were explored throughout the last two chapters and the both sides of the metaethical issue were given consideration on the matter. Kant's account of evil, though, has flexibility of its own and can solidify the internalist position on this matter rather than echo the same internalist talking points. The last section has opened up a possible gap in Kant's account that can perhaps make it conceptually possible for human beings to be evil for evil's sake. Paul Formosa and Irit Samet-Porat have suggested a similar view, that Kantians can maintain evil-qua-evil motivation as conceptually possible for human beings without undermining the significance of Kant's a priori principles. Their views will be helpful in

this section to establish another way of interpreting Kant's views on evil without outright rejecting his view of diabolism in *Religion*.

To continue from the previous section, there seems to be a discrepancy between the purely evil agent as described here and the diabolical agent as described by Kant in *Religion*. While both share the motive to do what is evil as evil, the diabolical agent renounces the moral law entirely as opposed to the purely evil human agent who is portrayed as still operating within the realm of morality that utilizes the moral law as inversely motivating. The purely evil agent here portrays a subtle dependency on the moral law's existence in their will as something perceptively revulsive and detestable. This dependency is due in large part to the mirror thesis which, outlined earlier in section 3.2, depicts the pure motive of respecting the moral law for its own sake as dependent upon an opposite capacity to act from a (pure) motive of hatred for the moral law. Paul Formosa also seems to implicitly embrace the mirror thesis below in defending the possibility of evil-qua-evil motivation for humans:

> "Just as we can act for the sake of the positive feeling of respect, so too we can act for the sake of the negative feeling of pain and humiliation…To act directly upon this hubristic hatred of the law is to choose evil qua evil. It is to choose evil immediately and not for the sake of (and indeed in spite of) any mediate interest or inclination that precedes the representation of the moral law."[192]

There is also a significant difference between the purely evil agent and Kant's wicked agent. As the previous sections indicated, wickedness is (self-deceptively) making evil out to be good in some perverse or twisted sense and not choosing evil as evil. That is, the wicked agent explicitly prioritizes the principle of self-love above the moral law but considers one's actions to be justifiably good. Using the mirror thesis, there should also be motivational poles to the irresistible "thrust" of the moral law.[193]

---

[192] Formosa, "Kant on the Limits of Human Evil", 197.

[193] Kant, *Religion*, 40 [AK 6:36].
As rational beings, Kant strongly believed that the moral law is an integral part of the practical reasoning of human beings. One way or another, there is no escaping the moral law's presence because (via the

Respect is merely one side of an agent's dispositional spectrum with revulsion/hatred at the opposing end. As the mirror thesis seemed to suggest earlier regarding opposite values, Formosa does not see any a priori reason from Kant's position that prevents human beings from feeling *only* respect for the moral law.[194] Furthermore, this attitude is motivationally distinct from wickedness in that the agent does not deceived himself in thinking what is evil is good. Samet-Porat also sees potential within this distinction, pointing out that "in such cases we are looking at a volitional structure in which goodness is recognized for what it is, but such a structure drives a person away by a deeply flawed rationality."[195]

Thus, between wickedness and diabolism, a conceptual gap is open for human beings to be purely evil in motive without trespassing on diabolical evil. Formosa further suggests that Kant could conceptually account for evil-qua-evil motivation as an extreme—and albeit rare—instance of human agency:

> "While spiteful hubris or perversity might be a sufficient ground to adopt maxims to pursue evil qua evil, this ability cannot turn us into devilish beings. A devilish being has an 'evil reason' and therefore lacks agency. A person who choose evil qua evil has agency and a will that, in the abstract, is good, but simply choose to make a particularly perverse use of that agency."[196]

Pure evil is not the same as diabolical evil because the moral law remains a motivational presence in the will's adoption of a principle of maxim-selection. While the moral law remains as an incentive (just completely inverted motivationally), the purely evil human agent cannot be relegated to one of Kant's three grades of evil. The agent is not self-deceived about the nature of their choice to do evil in the same way as the wicked agent. That is, within the conditions of

---

Reciprocity thesis) it is both necessary and sufficient for one's will to autonomously determine laws for oneself. Hence, this explains Kant's adamancy about the irresistibility of moral reasoning. Regardless of whether one conforms to the moral law or not, the moral law cannot be excised from human beings' freewill.

[194] Formosa, "Kant on the Limited of Human Evil," 198.

[195] Samet-Porat, "Satanic Motivations," 88.

[196] Formosa, "Kant on the Limits of Human Evil," 198.

frailty, impurity, and corruption, the purely evil human agent is not deceived about what one has chosen to incorporate as the basis of maxim-selection.

It was mentioned in section 1.2 that Kant's rejection of a diabolically human being relates to Anscombe and McNaughton's position that evil-qua-evil motivation is at bottom a closeted perverse evil that utilizes the term "evil" in an inverted-commas sense. Muchnik similarly describes Kant's rejection as not an expression of naiveté, but rather the belief "that finite rational agents, even as they disregard the moral law, act *sub specie boni*, no matter how distorted or perverse that supposition might be…we must represent our actions as pursuing something that matters to us, some good to which we bestow our interest."[197] Hence, one can take Kant's rejection as endorsing the view that any human case of evil-qua-evil motivation is merely sophisticated perversity or wickedness (i.e. evil as an instrumental value rather than intrinsically valuable for itself). Thus when Milton's Satan proclaimed "Evil, be thou my Good," Satan should not be thought as pursuing evil for itself but rather inventing his own conception of what is good to compete against God's notion of goodness.

The suggestion in section 1.2 to contest this view is that there is another sense one can value evil and not render it good in an evaluative sense. That is, evil agents can recognize and positively perform evil acts as evil without changing their views on what is good. Robert Dunn called it a formal sense of good based on a "criterion of success for any related action."[198] The proclamation "Evil, be thou my Good" could instead be read as revealing Satan's complete inversion of his motivational compass. As highlighted above, good no longer motivates in the sense that one respects the moral law but rather it inspires disgust and hatred. Just like the purely evil agent, Milton's Satan is not a wicked agent because there is a deliberate intentionality in choosing evil. He presumably knows what evil is according to moral discourse and seeks to

[197] Muchnik, *Kant's Theory of Evil*, 117.
[198] Dunn, "Is Satan a lover of the good?", 15.

perform it *because* it is evil. The fact that something is normatively good serves as motivation to do the contrary. To do evil for evil's sake is something that far exceeds the evil propensity in frail, impure, and wicked agents.

There are some problems with trying to fit evil-qua-evil motivation and the conceptual possibility of purely evil human agents within Kant's views. One problem is that self-deception plays a critical role for Kant in explaining how evil is imputable to judgments and actions. There is a general consensus and worry among Kantian scholars on this point.[199] If Kant's Reciprocity thesis hinges on viability of the moral law as one of the candidates for a will to govern itself, then how does self-deception apply to the purely evil human beings?

At face value, purely evil agency seems to rebuke this explanatory feature because evil-qua-evil motivation does not admit, within an agent's will to choose, self-deception. Samet-Porat remarks that such agents generally display a "surprising degree of self-reflection" about their own actions and perhaps have first-person knowledge of their own psychology and moral development.[200] Unlike other agents that can be discounted as ignorant (e.g. akratic, perversely evil, and amoral agents) or psychologically demented (e.g. psychopaths, serial killers, and wantons), purely evil agents exemplify the clarity and resolve to do evil within their motivational disposition (i.e. Gesinnung). How can pure evil be integrated into a Kantian framework but still distinguish itself from other (perverse) grades of evil? Perhaps there is a way to maintain Kant's view here, as well as his rejection of diabolism, and leave open the conceptual possibility of purely evil human beings.

[199] See Section 3.1 for more on Kant's explanation of self-deception with regard to the three grades of evil. For more information in general, Lawrence Pasternack's synopsis on Kant's self-deception is a great starting point in "Can Self-Deception Explain 'Akrasia' in Kant's Theory of Moral Agency?", *Southwest philosophy review* 15, no. 1 (1999): 91-93.
[200] Samet-Porat, "Satanic Motivations," 92.

Considering the fact that purely evil human beings are so different from the garden-variety evildoers one generally encounters in moral discourse, it should be expected that the basic understanding of self-deception needs to be adjusted. In moral discourse, many agents are not motivated to do evil at all. They sincerely want to do what is normatively good but may ignorantly do what is evil instead. Some agents have perverse motives and deceived themselves in adopting a twisted conception of good. Others can have amoral tendencies and convince themselves that their sincere moral judgments are not (inherently) motivating. Pure evil, being an anomaly among evildoers, has a different thought process and motivational structure than any standard agent.

Whereas most agents are victims to deceiving themselves about the knowledge of and judgment that "X is evil" due to frailty, impurity, or even corruption, the purely evil agent is not deceived in any of those ways about their choice to do evil. Formosa makes a similar point that "self-deception enters into the story only if [an agent] tries to justify [himself]."[201] Though purely evil humans do not attempt to explain away their evil actions, this should not exclude them from self-justification by some other avenue. By process of elimination, self-deception must occur in the purely evil agent's reversal of his motivational disposition of the moral law from respect to revulsion. The agent implicitly believes himself to be justified and that the inversion of his motivational compass towards evil is something that he desires for itself. As such, one can maintain Kant's view that self-deception accompanies immorality by suggesting purely evil agents deceive themselves in thinking that they want to do evil as evil.

Formosa in similar terms makes an interesting argument that people can be easily tricked into thinking someone is deliberately choosing evil. He utilizes Kant's account of passions in arguing that evil-qua-evil motivation can be mimicked in humans and that most cases of

[201] Formosa, "Kant on the Limits of Human Evil," 199.

destructive evil are instances of "passionate single-mindedness" or evil-qua-passion.[202] What makes this different from diabolism is that the human passions are the source of the self-deception for desiring evil for itself. It also poses a difficulty with this section's attempt to propose evil-qua-evil motivation as part of the evil propensity of human beings.

Whether or not this interpretation is successfully convincing, the reality of pure evil is another matter entirely—one that this inquiry cannot address at this time. While Formosa and Samet-Porat think pursuing evil for itself is conceptually possible for human beings and seek to constructively adjust Kant's views with their own, both are hesitant to speak beyond the coherency of the concept.[203] This difficulty was also mentioned in section 1.1 and 1.2 and must unfortunately be left without an answer. Nonetheless, people can find many things motivating in life. The prospect of being evil at face value is likely as much a motive as any other incentive for one's actions. However, there is no shame in acknowledging the rarity of evil-qua-evil motivation within human agency perhaps to the point of making it existent only as a conceptual possibility. Appeals to literature and other creative avenues are the best resources available as tools of analysis in this endeavor. But it would be a mistake to discount the plasticity of the human mind to motivate itself in unexpected ways, even a volitional/motivational structure that goes so far as to want to do what is evil for its own sake.

As discussed in section 3.2, pure evil should not be equated to diabolism. That is, evil-qua-evil motivation is not a sufficient condition for diabolism as Kant has presented it. To be diabolical additionally requires one to completely renounce the moral law as a viable basis for one's actions. But the purely evil human being, in pursuing evil for itself, depends upon the motivational inversion of the moral law. Rather than eliciting feelings of respect, the moral law

---

[202] Ibid., 205.
Formosa further remarks that "such agents can seem like devils who pursue 'intentional transgression' on 'principle'. However, passionate evil is not to be confused with evil qua evil…"
[203] See Samet-Porat, "Satanic Motivations," 84; and Formosa, "Kant on the Limits of Human Evil," 207.

inspires hatred and motivation to do contrary to what it dictates. This view of evil-qua-evil motivation preserves Kant's rejection of the diabolical and at the same time gives Kant's account much needed flexibility. Kant's views on evil are not being contested, but rather modified to include a higher grade of evil beyond or within wickedness itself.

Incorporating pure evil as an additional grade of evil within human beings, separate from the diabolical, allows Kantian reasoning to better accommodate the Dostoyevskian depths that humanity can descend. Since diabolism no longer has a monopoly on the motivation to do evil for evil's sake, it is not necessary to renounce the moral law in order to pursue evil as evil. In fact, the moral law is just as necessary for the evil will as it is for Kant's good will. The decisive factor is the volitional structure of the purely evil agent undergoing a complete reversal in motivations. Other aspects of Kant's account are maintained as well. Self-deception can still be attributable to immorality in the case of the purely evil agent, since it arises not in the choice itself but in the motivational shift of the agent. One can still agree with Kant that the diabolical is not humanly possible, but insist that a malignant practical reasoning can possibly surface through human self-deception of one's motivational allegiances as opposed to being mistaken about one's evaluative knowledge of good and evil (in the case of other evil agents). If Kant's three grades of evil are viewed as ways in which human beings can *motivationally* deviate from the moral law,[204] then the purely evil human being represents the true zenith of human evil—even if it only exists as a conceptual possibility.

There are less constructive views of Kant's account of evil. One way to address the predicament is to simply reject Kant's rejection. That is, not only should there be no a priori

[204] Kant, *Religion*, 35 [AK 6:32-33].
"The human being is *evil*, can signify nothing other than this: He is conscious of the moral law and yet has admitted the (occasional) deviation from it into his maxim." If one interprets the 'deviation' of human beings as scaled motivationally, then perhaps one can imagine a human being that is motivated by the ideal of diabolism. While this person would not be diabolical, his motivational dispositions would be comparable to a diabolical being—wanting to do what is evil because it is evil.

limitations on the freedom to choose evil for itself but also human beings should be able to will anything including renouncing the moral law and becoming diabolical. The previous view, inspired by Formosa and Samet-Porat, does not go this far; and it is unclear if negating Kant's position is necessary or even advantageous to the aims of this inquiry. Muchnik, among other Kantian scholars, has expressed strong reservations about completely doing away with Kant's stance on diabolism:

> "No matter how many examples we may present of individuals that systematically oppose the moral law in their conduct, Kant could always shrug them off: observable conduct does not decide upon the validity of the normative dimension of the will (Wille). From a purely normative perspective, the agent's opposition to the moral law is incapable of defining her freedom of choice (Willkur)."[205]

Kant's rejection of diabolical evil for human beings is more critical to his other projects than one might think. An implication of disregarding Kant on this issue may possibly result in the collapse of key a priori principles established earlier (e.g. the Reciprocity Thesis, Incorporation Thesis, etc). Unless one entertains the desire to rework Kant's internalism at its a priori roots, the subtle revisions here that incorporate evil-qua-evil motivation within human beings' evil propensity proves to be a simpler approach.

Hence, it seems much more viable to take the view that was put forward at the start. One can suggest, alongside Paul Formosa and Irit Samet-Porat, that Kant was perhaps too quick to dismiss evil-qua-evil motivation as a potentially human enterprise. One can affirm Kant's insistence that the moral law cannot be completely renounced (without thereby losing one's status as a free agent), but make the case that human beings as purely evil (i.e. evil-qua-evil motivated) can retain some remnant of the moral law when one utilizes—as Formosa emphasized earlier—a "hubristic hatred of the moral law"[206] as incentive to do the contrary. The error of the wicked agent is being self-deceived about the features of one's choice (i.e. making the principle of self-

[205] Muchnik, *Kant's Theory of Evil*, 113.
[206] Formosa, "Kant on the Limits of Human Evil," 197.

love to be good). The purely evil human being, though, does not try to justify evil as good but pursues evil precisely because it is contrary to the moral law. He principally deceives himself that his immediate hatred of the moral law is something worth pursuing as a formal good. Whereas the diabolical being renounces the moral law in its entirety and embodies a one-sided commitment to the principle of self-love, the purely evil agent has deceived himself in wishing to renounce the moral law when the very basis of his hatred of the moral law relies upon it. This view maintains Kant's rejection of diabolism for human beings, but brings much needed depth to evil that accommodates the potentiality of human beings reaching a quasi-diabolical state.

3.4 A Critical Objection

In light of this interpretation of pure evil, there is a serious objection that could possibly undermine the whole account. First, to briefly remind the reader, pure evil was first defined in this thesis as a purely principled stance to do what is evil primarily because it is evil. Whereas most, if not all, evil-doing is performed with some incentive in mind (e.g. power, pleasure, fame and glory, etc.), the incentive or motivation behind purely evil agency is the action itself. With Kant as the final component of this analysis, the previous section suggests that a purely evil human being is viable at the very core of one's motivational structure. Assuming that the mirror thesis is true, the moral law can be either a source of disgust and contempt or inspire respect and admiration. Hence, one can think of purely evil human beings as incorporating a motivational reversal of the moral law (i.e. that which is good) whereby evil is chosen precisely because it does not correspond with the moral law.

The objection one can bring to this interpretative solution is that it ultimately fails to specifically illustrate what is being defined. The agent in question is someone who has convinced himself motivationally that evil is worth pursuing for its own sake, but is the agent really purely evil or just another instance of perverse evil? Instead of a purely principled stance, the motivational reversal of the moral law may be the result of an ulterior motive such as the thrill (i.e. pleasure) of evil or some vain emulation of a fictional villain. For example, St. Augustine of Hippo's account in section 2.2 can be construed as evil-qua-pleasure or evil-qua-glory rather than evil-qua-evil. When Augustine said, "I had no motive for my wickedness except wickedness itself. It was foul and I loved it,"[207] the thrill of doing evil can count against the motive of doing evil for its own sake. Augustine's description of his childhood thievery seems to endorse pure evil, but other motives may be in play. Evil agents may be subtly motivated by other incentives.

[207] Augustine of Hippo, "The Depth of Vice: from *Confessions,*" 336.

Upon further reflection, this objection stems from earlier concerns in sections 1.1 and 1.2 about whether pure evil is, in reality, merely a sophisticated perverse evil. The iconic phrase "Evil, be thou my Good" from John Milton's *Paradise Lost* provokes different views on the mindset or motivational disposition of evil. The view endorsed by Elizabeth Anscombe and David McNaughton about Milton's Satan may be relevant to this objection. If bitterness, hatred, and contempt are basic attitudes of evil agents, then it naturally seems to follow that evil is a vain and rebellious redefinition of what is good. The negative reaction here seems to correspond to the objection that evil agents—whatever they may consciously affirm—mask their perversity in convincing ways. As such, the purely evil human being could be an agent that says their ground for action is purely principled but unconsciously have impure motives in play. But is it conceptually absurd for an evil-doer to have purely principled motivations in their actions? Must all evil mindset be supplemented by some underlying perverse incentive? At the very least, this thesis provided an alternative view for thinking otherwise.

Admittedly, this objection makes a fair point. The concern is that one simply does not know to what extent any agent (evil or not) is motivated to act. The conscious mind may itself mask other motives that are unknown even to the agent himself. In the case of Augustine, there is hardly any doubt that his childhood thieving elicited some kind of pleasure or thrill. Yet whether Augustine primarily stole those pears for the sake of stealing itself or for the thrill/pleasure of stealing is surely an open question. Furthermore, the mere existence of pleasure in evil-doing does not necessarily revoke the possibility of purely evil motivation for human beings. Just as moral saints may get pleasure out of principally being good for its own sake via supererogatory acts, evil agents may encounter other incentives that accompany the principled ground of their actions. To reiterate again what was said in chapter one, the difficulty of identifying pure evil in human terms should not be used as a verdict on its conceptual coherency. In light of this objection, the solution proposed in the previous section may be retained if amended that the

motivational inversion of the moral law does not prevent the interference of other incentives or circumstances. Human motivation is volatile and purely evil motivation can be overshadowed at a moment's notice.

CHAPTER IV

Conclusion: "What does it mean to be Evil?"

This thesis attempted to foster an inquiry that explored the motivational patterns or configurations of evil. These patterns are both subtle and nuanced, for one is dealing with the innermost workings of an agent's thought processes. Beginning from a conceptual standpoint, it was proposed in section 1.1 that evil agency can be divided between those who perform evil acts as a means to some other end and those who perform evil acts for their own sake (on presumably principled grounds). This initial distinction led to a discussion about metaethical theories concerning motivation in judgment—motivation internalism and externalism. Further, the addition of Kantian ethical theory elevates the tension that permeates these rival theories. There are several insights that one can take away from this thesis; a few will be explored below.

First, while the addition of Kant's account in *Religion* may be considered sudden and arbitrary, his approach and resultant views align closely with the aims of this thesis. To get to the core of the question "What does it mean to be evil?" it is critical to examine the preconditions that spur agents towards such acts. This requires looking at an agent's motivations and subsequent judgments on an a priori scale. As pointed out earlier in section 1.1, the danger of analyzing evil intentions on the level of appearances is the strong possibility of being misled by them. Experience only tells part of the story. It is within the domain of metaethics and other related fields (e.g. conceptual analysis, moral psychology, etc.) to reveal the underlying conflicts within agent deliberation.

This thesis labored to construct a conceptual view of evil that includes the prospect of positively choosing to do evil for itself—also called evil-qua-evil motivation. Sections 2.1 and 2.2 utilize the metaethical framework of motivation internalism (and externalism) to unravel the mechanisms in play when agents carry out evil acts. But a prevalent feature of many, if not all,

internalist views rejects the conceptual coherency of evil-qua-evil motivation. Socrates' implicit internalism in the *Meno* and *Republic* goes so far as to cast doubt on the existence of akrasia in moral agency—that a person can know what is morally required of them but be motivationally impotent to act accordingly. While tremendously influential and philosophically intriguing, Socratic internalism is largely outdated in relation to current philosophical literature. Plato's Socrates is unable to benefit from the added depth in human psychology starting with Freud and culminating in the ongoing discoveries of neuroscience. Human beings are much more complicated than previously thought.

Yet purely evil agency is still largely at odds with motivation internalism because of the tendency, first outlined in the contemporary views of Elizabeth Anscombe and David McNaughton, to conflate an agent's pursuits as indicative of what the person designates as a moral good. In other words, moral theorizing can take for granted that every agent desires or wants to achieve what he/she believes to be normatively good. It was granted that many evildoers do in fact have this sort of mindset whereby the term "evil" is used in an inverted-commas sense. These agents were referred to as "closet lovers of the good". To put it more simply, many evil acts are fueled by a twisted conception of what is good. Whether it is the desire for fame, fortune, or some other greater good, there is a supposed utility or benefit from the act that can explain an agent's motivation. The overwhelming majority of human evil falls under the label of perverse evil as it is defined in section 1.1. The critical question, though, that has dominated the bulk of this thesis is whether evil in some purely principled form is also possible for human beings.

Whenever moral discourse highlights agents like Milton's Satan that seemingly embrace evil for itself, there is an immediate tendency by many to see something incoherent with that position. But the fact that pure evil may be an anomaly that one can only find in literature or cinema is largely irrelevant to its conceptual coherency.

Kant's views in *Religion* offer a balanced approach that can supplement the conceptual framework of evil in this thesis. While Kant rejects the prospect of a thoroughly evil human being, his views at least leave open the possibility of beings that can possess such motivations (i.e. the devilish or diabolical will). His three grades of evil seem to give the internalist position more elegance and depth to the problems of akrasia and amorality. Furthermore, the interpretation offered in sections 3.2 and 3.3 could resolve the rough edges of Kant's views on diabolical evil and at the same time greatly alleviate the tension within motivation internalism.

Pure evil was initially defined in section 1.1 as "the performance of evil acts for their own sake". Over the course of this inquiry, its definition was enhanced through the discussion of de re/de dicto language use as well as the distinction between prudential and principled motives of morality in connection with Kant's hypothetical and categorical imperatives. The addition of motivation internalism and externalism put pure evil under scrutiny as to whether one could know and with sincerity affirm judgments about goodness without being motivated to act on them. Ultimately, with Kant guiding this framework, evil was redefined in a priori terms as when the will legislatively prioritizes one principle of maxim selection (self-love) over another (the moral law). As such, an agent's Gessinung or moral disposition to be good (or evil) is determined by whether the moral law is properly situated to determine one's basis for action.

Admittedly, the solution offered in section 3.3 does not make evil-qua-evil motivation any more likely. But the conceptual possibility and/or coherency of a principally evil human being can be preserved in Kantian terms if one considers the variations of human emotion and the effects it can have on moral motivation. Paul Formosa shares a similar view and also appeals to imagination and passion as an outlet for understanding evil-qua-evil.[208] With revulsion as the suggested feature of evil-qua-evil motivation, one can conceive a human being perhaps surpassing the garden-variety wickedness of evildoers when the moral law is used as an incentive

[208] Formosa, "Kant on the Limits of Human Evil," 207.

for self-love. Such an agent is not deceived about prioritizing self-love over the moral law and, hence, their will is not frail or impure. Nor is the agent redefining and justifying self-love as an evaluative good in place of the moral law—which falls under Kant's third grade wickedness and what this thesis calls perverse evil. However, the purely or principally evil human being falls short of diabolism, which renounces the moral law entirely, since the moral law as an incentive is inverted to affirm what is evil *because* it is evil. This solution preserves what Lawrence Pasternack calls "Kant's middle course between diabolism and unintentional immorality" since the moral law is not renounced and self-love (via sensible inclination) is not given precedence over reason.[209]

This leads to the most important insight of this thesis: motivation is not a one way street. Evil is not simple, especially when considering the conditions and circumstances that involve making judgments and acting on them. While there are general patterns of perversity in human evil, motivation are as flexible and quite adaptive to changes in both circumstance and agent mood. One should be aware of the intricacies of motivation and the impact that emotion can have on an agent's moral disposition(s). The details matter and are critical to fully explaining moral agency—good and evil. Kant's insistence that the moral law cannot be extirpated, only obscured through self-love, should not be ignored or dismissed. Though Kant's views on human freedom are limited to within the bounds of a structural heterogeneity between pure reason and inclination, there are still ways—if this interpretive solution is correct—in which evil motivations can subvert the moral law from within the darkest corners of man's passion. The potentiality for human beings to either heroically uphold duty to the moral law or catastrophically fall into the pit of self-love is endless, as Formosa rightly points out.

> "The depths to which humanity can sink are almost bottomless. But no matter how far we sink, as long as we retain our agency, and thus the capacity for pure reason to practically determine our actions, the hope for progress need not be

[209] Pasternack, "Can Self-Deception explain akrasia in Kant's theory of moral agency?", 93.

> completely lost. Humanity has it in it to approach the perfection of angels no less than the depravity of devils."[210]

Becoming familiar with the depths of evil at the fundamental level of motivation and judgment can better equip theorists in answering these and other issues of moral importance. In revealing humanity's potential devilishness, one can take the steps towards addressing it. This thesis has scratched the surface of these issues in the hope that future inquiries will yield an enormous treasure trove of knowledge and wisdom.

[210] Formosa, "Kant on the Limits of Human Evil," 207.

REFERENCES

Allison, Henry. *Kant's Theory of Freedom*. Cambridge: Cambridge University Press, 1993.

Anscombe, Elizabeth. *Intentions*. 2nd edition. Oxford: Blackwell, 1963.

Aristotle. "Nicomachean Ethics." In *The Basic Works of Aristotle,* edited by Richard McKeon, 935-1112. New York: Modern Library, 2001.

Augustine of Hippo. "The Depths of Vice: from *Confessions.*" Translated by John K. Ryan. *Vice and Virtue in Everyday Life* 8th ed. Edited by Christina Hoff Sommers| Fred Sommers, 333-336. Belmont: Wadsworth, 2010.

Ayer, A.J. *Language, Truth, and Logic*. New York: Dover Publications, 1952.

Barry, Peter Brian. "Moral Saints, Moral Monsters, and the Mirror Thesis." *American Philosophical Quarterly* 46, no. 2 (2009): 163-176.

Bensen-Cain, Rebecca. *The Socratic Method*. New York: Continuum International Publishing, 2007.

Blackburn, Simon. *Ruling Passions: A Theory of Practical Reason*. Oxford: Oxford University Press, 1998.

Dunn, Robert. "Is Satan a Lover of the Good." *Ratio (new series)* XIII, no. 1 (2000): 13-27.

Formosa, Paul. "Kant on the Limits of Human Evil." *Journal of Philosophical Research* 34, (2009): 189-214.

Fromm, Erich. *The Anatomy of Human Destructiveness*. Greenwich: Fawcett Crest, 1973.

Haybron, Daniel M. "Evil Characters." *American Philosophical Quarterly* 36, no. 2 (1999): 131-148.

James, William. "Will to Believe." In *Essays on Faith and Morals,* edited by Ralph Barton Perry, 32-62. Cleveland: Meridian Books, 1967.

Joyce, Richard. *The Myth of Morality*. New York: Cambridge University Press, 2001.

---. "Expressivism and Motivation Internalism". *Analysis* 62, no. 4 (2002): 336-344.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary J. Gregor. Cambridge: Cambridge University Press, 1998.

---. *Religion within the Bounds of Bare Reason*. Translated by Werner S. Pluhar. Indianapolis/Cambridge: Hackett Publishing, 2009.

Kauppinen, Antti. "Moral Internalism and the Brain." *Social Theory and Practice* 34, no. 1 (2008): 1-24.

Korsgaard, Christine M. *The Sources of Normativity*. Edited by Onora O'Neill. Cambridge: Cambridge University Press, 1996.

Lockie, Robert. "What's Wrong with Moral Internalism." *Ratio (new series)* XI, no. 1 (1998): 14-36.

Machiavelli, Niccolo. *The Prince*. Translated by George Bull. New York: Penguin Books, 2005.

Mackie, J.L. *Ethics: Inventing Right and Wrong*. New York: Penguin Books, 1977.

McGinn, Colin. *Ethics, Evil, and Fiction*. Oxford: Clarendon Press, 1997.

---. *Moral Literacy or How To Do The Right Thing*. Indianapolis/Cambridge: Hackett Publishing, 1992.

McNaughton, David. *Moral Visions*. Oxford: Basil Blackwell, 1988.

Melville, Herman. *Billy Budd*. Edited by F. Barron Freeman. London: Oxford University Press, 1948.

Mill, John Stuart. *Utilitarianism*. Edited by Oskar Piest. Indianapolis: Bobbs-Merrill, 1957.

Miller, Christian Basil. "Motivational Internalism." *Philosophical Studies* 139, (2008): 233-255.

Milton, John. *Paradise Lost*. Oxford: Oxford University Press, 2005.

Muchnik, Pablo. *Kant's Theory of Evil*. Lanham: Lexington Books, 2009.

Nietzsche, Friedrich. *Beyond Good and Evil*. Translated by Walter Kaufmann. New York: Vintage Books, 1989.

Pasternack, Lawrence. “Can Self-Deception explain akrasia in Kant’s theory of moral agency?”, *Southwest Philosophy Review* 15, no. 1 (1999): 87-97.

Plato. *Republic*. Translated by G.M.A. Grube. Indianapolis/Cambridge: Hackett Publishing, 1992.

---. “Meno.” *Five Dialogues*, translated by G.M.A. Grube, 58-92. Indianapolis/Cambridge: Hackett Publishing, 2002.

---. “Phaedo.” *Five Dialogues*, translated by G.M.A Grube, 93-154. Indianapolis/Cambridge: Hackett Publishing, 2002.

Railton, Peter. “Internalism for Externalists.” *Philosophical Issues* 19, no. 1 (2009):166-181.

Samet-Porat, Irit. “Satanic Motivations.” *The Journal of Value Inquiry* 41, (2007): 77-94.

Searle, John R. “How to Derive ‘Ought’ from ‘Is’.” *Philosophical Review* 73, no. 1 (1964): 43-58.

Shafer-Landau, Russ. “A Defense of Motivational Externalism.” *Philosophical Studies* 97, no. 3 (2000): 267-291.

Sneddon, Andrew. “Alternative Motivation: A New Challenge to Moral Judgment Internalism.” *Philosophical Explorations* 12, no. 1 (2009): 41-53.

Star Wars: Episode I – Phantom Menace. Directed by George Lucas. Marin County, CA: Lucas Films, 1999. DVD.

Star Wars: Episode III – Revenge of the Sith. Directed by George Lucas. Marin County, CA: Lucas Films, 2005. DVD.

Stocker, Michael. “Desiring the Bad: An Essay in Moral Psychology.” *The Journal of Philosophy* 76, no. 12 (1979): 738-753.

The Dark Knight. Directed by Christopher Nolan. Warner Brothers, 2008. DVD.

Tresan, Jon. “Metaethical Internalism: Another Neglected Distinction.” *The Journal of Ethics* 13, (2009): 51-72.

Watson, Gary. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.” In

Responsibility, Character, and the Emotions: New Essays in Moral Psychology, edited by Ferdinand Schoeman, 256-286. New York: Cambridge Press, 1988.

Wittgenstein, Ludwig. "Lectures of 1946-47." *Ludwig Wittgenstein: A Memoir*. Edited by Norman Malcolm. Oxford: Oxford University Press, 1966.

Williams, Bernard. "Internal and External Reasons". *Moral Luck*, 101-113. New York: Cambridge Press, 1986.

Zangwill, Nick. "Externalist Moral Motivation." *American Philosophical Quarterly* 40, no. 2 (2003): 143-154.

Zizek, Slavoj. *The Parallax View*. Cambridge: MIT Press, 2009.

VITA

Mark Smith Ferguson

Candidate for the Degree of

Master of Arts

Thesis: A PROLEGOMENON ON EVIL: "WHAT DOES IT MEAN TO BE EVIL?"

Major Field: Philosophy

Education:

Completed the requirements for the Bachelor of Arts in Philosophy at University of Central Oklahoma, Edmond, Oklahoma in 2009.

Completed the requirements for the Master of Arts in Philosophy at Oklahoma State University, Stillwater, Oklahoma in May 2013.