문학석사 학위논문

A Regulative Reading of Kant's Radical Evil

칸트의 근본악에 대한 규제적 해석

2019 년 8 월

서울대학교 대학원

철학과 서양철학전공

백 서 원

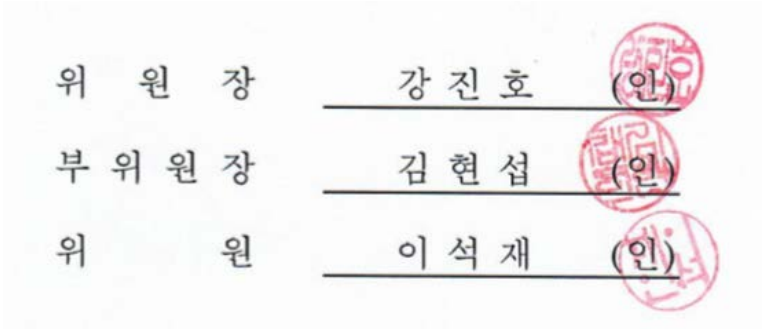
A Regulative Reading of Kant's Radical Evil

지도교수 김 현 섭

이 논문을 문학석사 학위논문으로 제출함
2019 년 4 월

서울대학교 대학원
철학과 서양철학전공
백 서 원

백서원의 문학석사 학위논문을 인준함
2019 년 8 월



Abstract

A Regulative Reading of Kant's Radical Evil

Seowon Baek
Western Philosophy Major
Department of Philosophy
The Graduate School
Seoul National University

This thesis is in league with the recent efforts to understand Kant's idea of radical evil through a coherent, reasoned perspective. To this end, I first give a preliminary account of Kant's theory of evil, and distinguish the three problems that most significantly obstruct a clear understanding of Kant's radical evil. These three problems together serve as adequacy conditions for a satisfactory interpretation of Kantian radical evil. In the second chapter, I examine two prior interpretations to evaluate their theoretical competency. As a result, I argue that since neither succeeds in giving adequate answers to the three problems, they both fail as acceptable accounts for radical evil. The majority of this thesis is focused on the third chapter, which is mainly an attempt to endorse and consolidate an alternative way of understanding radical evil. I argue that Kant's thesis of radical

evil is better understood as a postulate, or necessary hypothesis, that serves a regulative role for the Kantian moral discipline, one which requires that the agent *presuppose*, rather than identify as a universal matter of fact, that oneself has a deep-seated propensity to evil. I aim to contribute to this regulative reading of radical evil by (i) inspecting Kant's usage of vocabulary; (ii) strengthening and amplifying the core arguments for the regulative view; (iii) presenting an additional, practical reason to favor the regulative reading over others; and (iv) identifying and refuting a possible objection against the regulative reading.

Keywords: Immanuel Kant, radical evil, moral disposition, supreme maxim, inscrutability thesis, regulative, presupposition, self-deception, moral growth, moral self-perfection.

Student Number: 2016-20081

Table of Contents

Abstract (English) i

Introduction 1

1. Kant's Radical Evil: The Essentials 4

1. 1. The possibility of evil 4

1. 2. The supreme maxim and radical evil 8

1. 3. Moral regeneration 13

1. 4. The three interpretive puzzles 16

2. Prior Interpretations 19

2. 1. Allen Wood's anthropological reading 19

2. 2. Problems with the anthropological reading 20

2. 3. Henry Allison's deductive reading 23

2. 4. Problems with the deductive reading 25

3. The Regulative Reading 26

3. 1. Motive and textual evidence 26

3. 2. Presupposition and belief 36

3. 3. Markus Kohl's moral ascetic argument 44

3. 4. Evil as self-deception 46

3. 5. Merits and practical implications 48

3. 6. The cancer argument 54

Conclusion 59

Bibliography 61

Abstract (Korean) 63

Introduction

In *Religion Within the Boundaries of Mere Reason* (1793), Immanuel Kant presents and develops the thesis that "the human being is by nature evil" (*Religion*, 6:32). This claim, also known as Kant's doctrine of radical evil, has perplexed many readers, in Kant's own time and today alike. Those who had regarded Kant as a prominent exponent of the Enlightenment were disappointed, as they considered the doctrine of radical evil to be Kant's compromise with the Church in presenting a philosophical rendition of the Christian doctrine of original sin. Critics rebuked Kant's conception of evil, some condemning it as "morally perverse"[1] and others dismissing the notion as part of Kant's "speculative essays" about human history which should not be taken as seriously as his critical works.[2] Even the more charitable readers altogether rejected his idea, protesting that Kant's outlook on evil, by ascribing the "propensity to evil" as a universal property to all finite rational beings, trivializes the concept and lacks the explanatory power required to account for the extraordinary atrocities that appalled the world and properly deserve to be called evil.

There are a number of reasons why Kant's theory of evil elicited such unfavorable responses, but perhaps the most conspicuous one is that while Kant makes such a sweeping claim about human nature, his reasoning is not as clear. There are apparent inconsistencies in the text where Kant seems to contradict himself both within the book and beyond, against prior commitments in his overall

[1] See Richard J. Bernstein, "Radical Evil: Kant at War with Himself," 71.
[2] See Daniel O'Connor, "Good and Evil Disposition," 298.

system of practical theory. It is only since the early 2000s that significant discussions began among those who attempt to resolve these alleged inconsistencies and make sense of Kant's idea, seeing that what at first glance seems to be a wildly drastic and even misanthropic claim may after all provide meaningful insight about human nature and the quest for moral maturity. This thesis stems from the same motivation, and intends mainly to explicate Kant's idea of radical evil in a charitable light. Accordingly, in the first chapter I will offer a succinct explanation of Kant's theory of radical evil to illustrate its elemental aspects.

Meanwhile, there has been a general lack of clarity regarding which issues and themes of *Religion* should be considered central to the doctrine of radical evil, and it has been the frequent case for commentators to focus only on partially resolving the tensions that appear in *Religion*, or to spiral into an extraneous debate over different exegetical perspectives. The main objective of this thesis is neither to embark on an exegetical task of Kant's various texts, nor to dissect the intricate structure of Kant's system of moral philosophy in detail. Rather, the first contributive aim of this thesis is to build a structured criterion to analyze Kant's radical evil. This will take the form of three interpretive puzzles that arise in *Religion*, and whether these puzzles can be successfully resolved will serve as a standard to evaluate the different interpretations of Kant's radical evil. Subsequently, in the second chapter I will examine two major interpretations of Kant's radical evil to see how they fare against the three puzzles. The second chapter will cover in turn the "anthropological" reading suggested by Allen Wood, and the "deductive" reading as proposed by Henry Allison. I will argue that each of

these interpretations fails to resolve at least one puzzle, and faces further problems of its own.

The third and principal aim of this thesis is to endorse one possible way of understanding Kant's notion of evil, and the insights it can provide into the conditions of moral discipline. Therefore, in the third chapter, I will introduce an alternative interpretation of Kant's radical evil, namely, the "regulative" reading. The major motivation for the regulative reading is that it is more appropriate to understand radical evil as a postulate, or necessary hypothesis, that serves a regulative role for the Kantian moral discipline, one which requires that the agent *presuppose*, rather than identify as a universal matter of fact, that oneself has a deep-seated propensity to evil. Based on an initial proposal of the regulative reading by Markus Kohl, I will illustrate how the regulative reading not only can provide solutions to all of the three interpretive puzzles, but also carries an additional theoretical benefit in portraying how the agent's assumption of radical evil plays a crucial part in the moral regeneration as conceived by Kant, which can even be sanctioned from a contemporary perspective regarding moral character. Lastly, I will discuss a possible objection against the regulative reading, and see if it can be refuted.

1. Kant's Radical Evil: The Essentials

1.1. The possibility of evil

It helps to take note that one of Kant's major motives to define human evil is to explain its imputability to the agent; evil is what we can blame the agent for. It is what the agent ought to have resisted or abstained from, but nonetheless chose not to. We are familiar with the Kantian deontological principle that "ought" implies "can." Hence, for Kant, it is important to rationalize how evil is "freely" chosen by the agent, for otherwise the agent may be exculpated from the blame. To this end, Kant's major claims about human evil consist of several themes that connect human agency with moral evil. At first glance, the concept of evil seems not so easy to square with Kant's overall practical theory, in which morality is itself closely connected to human will and freedom. In the *Groundwork of the Metaphysics of Morals,* Kant specifies as follows:

> [W]hat, then, can freedom of the will be other than autonomy, that is, the will's property of being a law to itself? But the proposition, the will is in all its actions a law to itself, indicates only the principle, to act on no other maxim than that which can also have as object itself as a universal law. This, however, is precisely the formula of the categorical imperative and is the principle of morality; hence a free will and a will under moral laws are one and the same. [*Groundwork,* 4:447]

As can be seen from the above excerpt, Kant identifies the human free will to the will under moral law. On the other hand, an action that does not conform to the universal lawgiving form is seen to be dependent on the natural law of causality, as Kant describes in the *Critique of Practical Reason*:

> [I]f no determining ground of the will other than that universal lawgiving form can serve as a law for it, such a will must be thought as altogether independent of the natural law of appearances in their relations to one another, namely the law of causality. But such independence is called *freedom* in the strictest, that is, in the transcendental, sense. Therefore, a will for which the mere lawgiving form of a maxim can alone serve as a law is a free will. [*CPrR*, 5:29]

From Kant's perspective, an action is rational only insofar as it conforms to the moral law, and is therefore free and autonomous, while an immoral action is by definition an "unfree" or "heteronomous" one. This leads to the infamous objection raised by Sidgwick; according to Kant's practical philosophy, an immoral action is never freely chosen by the agent in the sense that it has been determined by the natural law of causality. If this is so, agents cannot be blamed with their immoral actions, and therefore an immoral action that can be imputed to the agent is logically impossible for Kant.

Since in *Religion* Kant delves right into the topic of human immorality, the discussion of which has been somewhat sparse in his earlier works of practical philosophy, a suitable theoretical device is required to show how human evil is possible in a way that it can be held responsible by the agent. For this, Kant employs two stratagems. Firstly, unlike in his preceding works, Kant questions the morality of agents, not of actions. In other words, the proper topic of *Religion* is to give an account of moral character—what it means for a *person,* not an *act,* to be morally good or bad. This allows room for the discussion of how human agents with free will can choose to act in a way that does not conform to the moral law, because even though an immoral action is itself a result of the law of causality and cannot be blamed upon, if the agent who has somehow *chosen* to act it out has

done so through a practical decision-making process, the immorality may still be imputed to the agent.

To make this process possible, Kant makes extensive use of the distinction between human *Wille* and *Willkür*. In *Religion*, Kant depicts the human faculty of volition to be construed of two distinct, but unified parts: the will (*Wille*) and the power of choice (*Willkür*). The difference between the two is that the former serves as a "lawmaker" that legislates norms by practical reasoning, while the latter is a faculty of "execution" that makes the executive choices to adopt the legislated maxims. The crux is that while *Wille* is to be identified with practical reason as we are already familiar with, *Willkür* can be affected by incentives other than the respect for the moral law. So even though *Wille* can only but legislate norms that conform to the moral law, it is the job of *Willkür* to choose to act upon those norms, or upon other maxims that incorporate incentives apart from the moral law, which would render the choice *immoral*. If our power of choice makes the executive decision to act upon maxims legislated by incorporating respect for the moral law, then we would have made an *autonomous* choice. If, on the other hand, we choose to act upon maxims that do not comply with the categorical imperative, we are making a *heteronomous* choice; however, in both cases the agent can be held responsible for the choice, because they were both *spontaneously* made through the unified faculty of volition. In this way, it becomes technically possible for an agent to be held accountable for the non-autonomous choices that one makes, because the choice is still made spontaneously, and in a sense, "freely."[3]

[3] Part of this explanation is owed to Paul Formosa (2007).

Throughout *Religion,* Kant uses the word *Willkür* to describe the responsibility of the agent's decision-making, and therefore of the agent's immorality. He proposes what is known as the "incorporation thesis," which states that the "freedom of the power of choice has the characteristic […] that it cannot be determined to action through any incentive *except so far as the human being has incorporated it into his maxim* (has made it into a universal rule for himself, according to which he wills to conduct himself)" (*Religion,* 6:23-24), which makes it clear that even though *Willkür* is not immune to other incentives, it is still the agent's responsibility to choose to act according to the moral law, and abstain from responding to other incentives. Kant stresses this accountability as follows:

> [H]ence the action can and must always be judged as an *original* exercise of his power of choice. He should have refrained from it, whatever his temporal circumstances and entanglements; for through no cause in the world can he cease to be a free agent. It is indeed rightly said that to the human being are also imputed the *consequences* originating from his previous free but lawless actions. [*Religion,* 6:41]

At this point, one may ask whether the morality of an agent is determined with each and every choice, so that a human being can be at a time autonomous and at others heteronomous, or in other words at a time morally good and at others morally evil. Kant denies this to uphold a claim that he calls "moral rigorism," meaning that human morality involves an excluded middle; no one is both morally good and evil at the same time, or neither morally good nor evil—it is always one or the other. The reasons for this claim will be examined in the following section.

1.2. The supreme maxim and radical evil

It is well known that according to Kant, human beings act upon maxims. Maxims are principles that rationalize the connection between action and intention. If we act according to some maxim, we aim to get to an end through a means. Upon closer inspection of the process of our practical deliberation, we find that maxims are multilayered, in the sense that some maxims "ground" others by providing rational justification for them. For instance, when I spend money on warm clothing, my friend may ask why I am choosing to spend that money on clothing instead of on other commodities. Then I would answer, because I have chosen to act upon the maxim to stay as warm as possible this winter. My friend may then ask why I have chosen that maxim, and I'd say that it is because I have chosen to act upon the maxim to be comfortable and avoid possible illnesses from the cold. If I am further probed with the same question, my reply would be that I have chosen to act upon the maxim to live a safe and happy life. We can see a hierarchical relationship forming between these maxims. The maxim on the surface level of individual decisions is justified by the maxim *beneath* it, and in turn that maxim is grounded by another maxim below. If we trace the connection of our maxims this way down to the deepest level, Kant claims that we will find a single "ultimate subjective ground of the adoption of maxims," one that serves as a basis to justify all other maxims. This ultimate ground, the "maxim of maxims," is what Kant calls moral disposition, or *Gesinnung*. Since this disposition would be the fundamental tendency of the agent in situations of moral choice, it is the criterion by which an agent's moral condition can be evaluated. In other words, one is morally good if one has a good moral disposition, and evil if one has an evil moral

disposition. The concept of *Gesinnung* is central to Kant's theory of moral character, because it is due to this concept of disposition that we can interrelate the individual, isolated actions of agents and understand them as manifestations of an agent's underlying moral character.

As to define what "good" and "evil" moral dispositions are, Kant contrasts the two ways that human beings can choose to incorporate an incentive into their maxims through their power of choice. A person may choose to adopt maxims to act upon, and only upon, respect for the moral law—that is, only in compliance with the categorical imperative. Otherwise, one may choose to adopt maxims to act primarily upon self-love; to prioritize one's self-interest over universalized maxims, and comply to the moral law only when it does not conflict with one's own benefit. These two incentives of the respect for the moral law and of self-interest are permanent appeals to our *Willkür*, and cannot be eradicated or incapacitated with regards to their appeals to our power of choice. Since the two incentives are configured in a way that one is exclusive of the other, and the structure of the hierarchy of maxims is never arbitrary or ambiguous but can only be rationally grounded by a single choice of incorporating one incentive or the other, these two incentives are seen to be mutually exclusive and collectively exhaustive. This is the reason that Kant rejects the possibility that a person's morality can oscillate by every individual choice that one makes, a claim known as the "rigorism thesis," which states that an agent is always either morally good or morally evil, and never both or neither. Consequently, moral goodness or evilness is a matter of subordination of incentives: "*which of the two he makes the condition of the other*." (*Religion*, 6:36)

For Kant, therefore, a person is not morally good on account of one's good *deeds*, but of good *intent*, and not just on a shallow level, but on the deepest level of commitment that lies beneath all individual everyday decision-makings. Recall that in the *Groundwork*, Kant accuses the shopkeeper who does not overcharge inexperienced customers out of concern for his own reputation and benefit to have acted out of mere self-love. Said shopkeeper has chosen to comply to the moral law only insofar as it does not disrupt his self-interest; it was only by chance that his actions that were intended for his own benefits corresponded to the moral law, and therefore this shopkeeper has an evil disposition. Only agents who have chosen respect for the moral law as "their sole and supreme incentive" are considered morally good—even an occasional deviation is inexcusable. In an extremely hypothetical case, it is possible for a person to have lived an entire life acting in compliance with the law, and still be morally evil at heart, because all those actions were compatible with the law only by accident, while the true intent, unbeknownst even to oneself, was that of self-love.

At this point, it seems that Kant is placing too high a standard for moral goodness, on an impossible level to achieve, a ruthlessly idealistic demand for perfection. What then is the point in trying to do good and be a good person, if a single mistake, a single prioritization of self-interest, betrays immorality? This is in fact an apt question to ask, as we will see later that this idea of moral perfection plays a central role in Kant's theory of moral discipline. Meanwhile, Kant does not say that there is no difference between a person who knowingly defies the moral law and another who strives to do good but makes intermittent "mistakes." He portrays a threefold manifestation of evil, in proportion to its degree. The first and

weakest grade of the manifestation of evil is called the *frailty* of human nature, and it refers to the state where one is aware of what one ideally ought to do, but when it comes to actual implementation, one's moral commitment is too weak to overcome other interests. It also includes cases where one knows what is *generally* right, but makes exceptions for oneself by attempting to justify one's actions through appealing to emotions or circumstances. The second grade of *impurity* is the state of mixed motives, where the agent acts partly due to moral law and partly due to self-love. For instance, I might give to charity mostly because I know that it is the right thing to do, but also because it makes me feel good about myself, although I might not admit the latter motive. Notice that these two grades of manifestation can be quite commonly found, and that in both cases some form of self-deception plays a part, either by convincing oneself that a moral exception can be made, or by concealing a mixture of motives to pretend one is morally righteous. The third and worst grade of human evil is called the *depravity, corruption,* and *perversity* (all used interchangeably) of human nature, which refers to the state where an agent intentionally and consciously subordinates the incentive of the moral law under that of self-love. One may choose to deceive, violate, invade, and trample over others knowingly, in order to gain personal success or pleasure. This may happen less frequently than the former two, but often enough for us to witness on the media. People who disregard others as equals, considering others as instruments and objects for their own interest, whether from a criminal motive or a highly manipulative one, have always plagued the human society. Seen this way, it seems quite the rare case that we do good for the right reasons. In fact, it is Kant's claim that a morally good person is not just scarce, but there is simply *no one* who is

morally good, and that everyone initially has reverted the two incentives to prioritize self-love.

So far, I have briefly summarized Kant's major claims regarding human evil. For the sake of convenience, here I will give a paraphrase of Kant's thesis of radical evil in a way that it incorporates the essential claims discussed above, formulated as follows:

> *Thesis of Radical Evil*:
>
> Every human being has freely chosen an **evil disposition** (or, **propensity to evil**) which is the agent's supreme maxim, or the ultimate subjective ground for all other maxims, to subordinate the incentive of the moral law under that of self-love.[4]

In *Religion*, Kant seems to be asserting that this evil disposition is *universal* in the sense that no human being is exempt from it; furthermore, it is *inextirpable*, and so deep-seated in humanity that it is enough to be called a *natural* propensity to evil. Therefore, for Kant, the evil in human being is *radical* not in the sense that it is extreme, but it is so deeply embedded in human nature that it corrupts our ability of choice at its very root. Yet he also stresses throughout the entirety of his discussion that this evil is nonetheless *imputable* to us because human beings must be held accountable for their own evil maxims.

[4] Here I will follow the convention to identify moral disposition(*Gesinnung*) and the propensity to evil(*propensio*). The reformulation of the thesis of radical evil as suggested here is based on Kant's own discussion in *Religion (*6:32-37).

1.3. Moral regeneration

If humans are thoroughly and irrevocably evil as Kant suggests, are we all doomed to sprawl in our own filth? Kant does not leave matters that way; after depicting a bleak picture of the human moral condition, Kant describes a process of moral reform through which we can shun our innate evil and eventually become a good human being. In order to understand this process, we need to first look at what Kant says about virtue. According to Kant in the *Metaphysics of Morals*, complete or perfect virtue is a near-impossible ideal for us (MM 6:409); it requires both the agent's adoption of good maxims on (and only on) the basis of the incentive provided by respect for the moral law, and a "strength of will" that enables agents to exercise such maxims with unwavering stability (MM 6:405). The ideal virtue being such, it is possible (and likely in most cases) for agents to adopt good maxims from the right incentive and even act upon them, though they may still falter in the process and occasionally deviate or "relapse" by making the wrong choices or having mixed motives. Kant acknowledges that finite agents are always vulnerable to temptations to deviate from moral laws.

Another important point to be noted about Kant's idea of moral character is that moral development *cannot* be a matter of mechanical habituation that one acquires through aligning one's actions according to the moral law without a fundamental change in the hierarchical structure of maxims, because that would be incompatible with the idea that moral agency is free. Kant denounces such idea of virtue in the following excerpt:

> Virtue here has the abiding maxim of *lawful* actions, no matter whence one draws the incentives that the power of choice needs for such actions. Virtue, in this sense, is accordingly acquired *little by little*, and to some it means a long habituation (in the observance of the law), in virtue of which a human being, through gradual reformation of conduct and consolidation of his maxims, passes from a propensity to vice to its opposite. But not the slightest *change of heart* is necessary for this; only a change of *mores*. A human being here considers himself virtuous whenever he feels himself stable in his maxims of observance to duty─though not by virtue of the supreme ground of all maxims, namely duty, but [as when], for instance, an immoderate human being converts to moderation for the sake of health; a liar to truth for the sake of reputation; an unjust human being to civic righteousness for the sake of peace or profit, etc., all in conformity with the prized principle of happiness. [*Religion,* 6:47]

Therefore, for Kant, a truly virtuous person is one who, in every situation of moral choice or conflict, performs free and autonomous moral deliberation, spontaneously chooses to incorporate only the incentive of the respect for the moral law into one's maxims, and acts upon it. This also indicates that the struggle to silence the temptation of the incentive of self-love is incessant and ever-present in situations of moral choice (MM 6:409).

Since a mere habituation of actions that abide by the law is not sufficient to make an agent morally good, Kant proposes a two-stage model of moral reform, that "a revolution is necessary in the mode of thought (*Denkungsart*) but a gradual reformation in the mode of sense (*Sinnesart*)" (*Religion,* 6:47). In other words, Kant claims that "a human being's moral education must begin, not with an improvement of mores, but with the transformation of his attitude of mind and the establishment of a character, although it is customary to proceed otherwise and to fight vices individually, while leaving their universal root undisturbed." (*Religion,* 6:48)

The first stage of this moral reformation, a revolution in the mode of thought, consists of a "single and unalterable decision" of an agent to restore the reversed subordination of incentives in his supreme maxim. This revolutionary transition, however, happens not by an external event or stimulus (because then it will be a result of the natural law of causality and hence cannot be "free"), but by the slow but ceaseless precipitation of the quiet voice of practical reason (*Wille*) that has been urging us from within to heed to the authority of the moral law and recover the respect for it as the proper supreme maxim. This is to establish the good disposition, in which the respect for the moral law is incorporated as a self-*sufficient* incentive into the power of choice, without being conditioned by other incentives. This recovery makes the agent "receptive" to the good, but not quite yet a good human being. The second stage of Kant's moral regeneration requires that the agent engage in an "incessant laboring" to stay true to this newly adopted disposition, though it is common for the agent to falter in this path and make occasional lapses as seen earlier by Kant's description of virtue. As Kant describes, "change is to be regarded only as an ever-continuing striving for the better." This is a very slow and gradual process, and as we will see later this is not a visible process either. Although the thesis of radical evil and this process of moral regeneration are closely interrelated, attempts to understand the two aspects of Kant's theory in unison have been scarce. In the next section, I will specify the issues that arise in comprehending Kant's theory of evil which must be resolved if a satisfactory interpretation of the theory is to be offered.

1.4. The three interpretive problems

Kant's discussion of evil, though its central thesis is asserted with surprising force, is not quite ample at length nor in detail, which makes it even more difficult to comprehend. So far, there has been various attempts to make sense of Kant's doctrine of evil, but the overall discussion lacked a coherent acknowledgment of the problems regarding the interpretation of radical evil. Due to such lack, the debates tended to collapse into a chronic tussle between different exegetical views which led to no general consensus. Extracting the most prominent issues from prior works in the literature, here I identify three major aspects of Kant's theory of evil that impede a consistent understanding of it. I do not claim originality to any of these problems, as they have been vastly proposed in different works of many other Kant scholars;[5] however, as there hasn't yet been a proper establishment of adequacy conditions to evaluate the various interpretative efforts at Kant's radical evil, I propose that these three problems function as evaluative standards in deciding which interpretation gains the upper hand.

The first, most well-known problem is that while Kant contends that every human being is evil without exception, he does not provide an official proof to back up this claim. He merely writes, "we can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being, in view of the multitude of woeful examples that the experience of human *deeds* parades before us." (*Religion*, 6:33) This report is instantly perplexing, for it appears evident that

[5] See O'Connor (1985), Wood (1999), Bernstein (2001), Allison (2002), and Formosa (2007), among many others.

empirical examples, however in multitude or woeful they may be, are not sufficient to support such a sweeping generalization, not to mention the apparent incongruence with Kant's meticulous attempts for a transcendental inspection of the human nature and capabilities that we are familiar with throughout his critical works. Surely Kant of all people would know that enumerative induction is no way to prove a judgment that could even be seen as a synthetic *a priori* judgment.[6]

The second problem issues when Kant denies that evil is a natural property of humans because evil should be something that a human being can be held accountable for, but right afterwards depicts evil as *entwined with*, or *rooted in*, human nature. See the below excerpt:

> Now, since this propensity must itself be considered morally evil, hence not a natural predisposition but something that a human being can be held accountable for, and consequently must consist in maxims of the power of choice contrary to the law and yet, because of freedom, such maxims must be viewed as accidental, a circumstance that would not square with the universality of the evil at issue unless their supreme subjective ground were not in all cases somehow entwined with humanity itself and, as it were, rooted in it: so we can call this ground a natural propensity to evil, and, since it must nevertheless always come about through one's own fault, we can further even call it a *radical* innate *evil* in human nature (not any the less brought upon us by ourselves). [*Religion*, 6:32]

It is at once confusing to figure out what exactly Kant intends here; he seems to be saying that evil is an innate, natural property, and yet at the same time it is

[6] There are dissenting opinions on whether the thesis of radical evil is a synthetic *a priori* judgment or not. Paul Formosa (2007) explicitly endorses that it is, and also Henry Allison (2002); but other critics such as Wood or Grimm who proffer different interpretations would probably disagree. Though it is not a matter of importance in my discussions, I tend to think that it is. However, as I will discuss later, as far as the regulative reading is concerned, it matters not as much whether the thesis is synthetic *a priori* or not, because even if it is, Kant is not trying to declare this judgment to be actually true; one need only *presume* it to be.

something that we brought upon ourselves as free agents and therefore not any less blamable. How a coherent explanation of this can be possible is the second crucial puzzle to solve in Kant's theory of radical evil.

Finally, Kant stresses several times in *Religion* that evil is something *inextirpable*, meaning it is not possible for human individuals to eradicate their evil inclination. However, Kant also states that it is a duty for individuals to become morally better persons, which in Kantian framework implies that humans are indeed capable of doing so; Kant describes a process of "conversion," through which only can humans break free from their inextricable evil. As examined earlier, this process takes on a two-stage model; it begins by a sudden revolutionary determination to commit entirely to the moral law, and then is followed by incessant efforts to change one's conducts to conform to the law, not only on the level of observable actions, but also on the internal, dispositional level. The third problem in interpreting Kant's idea of evil is how human evil is thought of as fundamentally inescapable, yet also it is required of individual persons to become morally better, and human evil is considered ultimately purgeable.

These three tensions are tasks for every attempt at a charitable reading of Kant's doctrine of radical evil to resolve. Whether these puzzles are solved so that Kant's radical evil can be explained consistently will serve as a standard to evaluate the attempts to interpret Kant's radical evil. Therefore, a plausible account of Kantian evil would be one that is able to present an interpretation that coherently conciliates the above three issues. In addition, I propose that it would be a further theoretical advantage if the interpretation can also provide some insights into the practical

implication of such a conception of evil; what Kant's motive was in depicting human evil in such a radical way, and why it is meaningful to understand human evil as Kant proposes. In what follows, I will briefly examine two major attempts at interpreting radical evil, and identify that they both fail to solve the above three problems.

2. Prior Interpretations

2.1. Allen Wood's "anthropological reading"

One of the earliest and most influential interpretations of Kant's doctrine of radical evil is suggested by Sharon Anderson-Gold[7] and developed by Allen Wood[8]. Here I will focus on Wood's argument, which is an attempt at justifying the propensity to evil by appealing to the anthropological conditions of human beings, and hence understanding Kant's doctrine by placing it within the socio-teleological context of the Kantian system. Wood identifies radical evil with "unsocial sociability," which can be roughly understood as a characteristic of our natural predisposition (*Anlage*) to humanity that drives us to make comparisons of ourselves against others and strive to attain self-worth and happiness through competition. This natural predisposition is what propels the scientific, political, and

[7] See Sharon Anderson-Gold, "God and Community: An Inquiry into the Religious Implications of the Highest Good."
[8] See Allen Wood, *Kant's Ethical Thought*, and "Kant and the Intelligibility of Evil."

cultural progress of the human civilization in the macroscopic, teleological viewpoint of human history, but when misused or corrupted it leads to social vices such as "envy, ingratitude, joy in others' misfortunes," and other such vices that Kant condemns as "diabolical." Based on some textual evidence, Wood argues that "Kant explicitly attributes the corruption of human nature to the *social* condition of human beings," and further claims that "the anthropological reading of the doctrine of radical evil also implies that evil has its *source* in social comparisons and antagonisms."[9] This, according to Wood, explains why the world is teeming with immorality: our propensity to evil originates if, and *only if*, humans live in proximity to each other. Given the developed state of human civilization and our unsocial sociability, humans are vulnerable to antagonizing each other and hence prone to the perpetration of evil acts even as extreme as terrorism and war.

2.2. Problems with the anthropological reading

One imminent problem with the anthropological reading is that regarding unsocial sociability as equivalent to radical evil yields consequences that contradict Kant's own major claims. There is in fact a passage in *Religion* which Wood uses to back his claim, where Kant declares that "[e]nvy, addiction to power, avarice, and the malignant inclinations associated with these, assail his nature, […] *as soon as he is among human beings* […] it suffices that they are there, that they surround

[9] See Allen Wood, "The Radical Evil in Human Nature" in his *Kant's Ethical Thought*, 287-290.

him, and that they are human beings, and they will mutually corrupt each other's moral disposition and make one another evil" (*Religion,* 6:94). However, to read this passage as a strong claim as Wood does, is to regard Kant as if he is insisting that the human social condition is a *necessary and sufficient* condition, itself enough to determine, an evil disposition. This, however, is clearly contrary to Kant's intention; if human evil originates from social condition, and it is normally the case that no human being is born and raised outside a social structure of some sort (one might here point to a possible "isolated man," but to say humans can only obtain a good disposition in isolation also involves a misrepresentation of Kant's views on humanity), how are we to avoid the consequence that the source of evil is already and inescapably determined for humans? How are we, then, to square this claim with Kant's lucid declaration that evil disposition is an independent *act* of the individual agent's free will? Although Wood distinguishes the *source* of radical evil from the grounds for our *moral responsibility*,[10] he provides no further argument or reason to convince that the two should be distinct. It is more plausible to read the above passage as an exaggeration and understand Kant to be stressing the grave difficulty that the presence of other people impose upon us in retaining moral integrity. Therefore, the anthropological reading may explain the first of the three problems by attributing the universality of human evil to the unavoidable social conditions of humans, but it fails to explain the second and third problem, because there is no viable solution to the tension between the innateness of evil and the individual agent's free act to choose evil, and also because Wood entirely omits

[10] Wood, "The Radical Evil in Human Nature," 289.

how radical evil seen as unsocial sociability is related to the Kantian moral conversion.

I would like to briskly point out one more problem with the anthropological reading. Kant initially views the three natural predispositions as themselves "not only (negatively) *good* (they do not resist the moral law) but they are also predispositions *to the good* (they demand compliance with it)" (*Religion,* 6:28). On each of these predispositions "can be grafted all sorts of vices (which, however, do not of themselves issue from this predisposition as a root)" (*Religion,* 6:26-27), and such vices, according to Kant, are the results of a *corruption* of the predispositions. But what, then, causes this corruption? Since Wood equates unsocial sociability with radical evil, he would have to answer that our unsocial sociability itself lends us vulnerable to the corruption, but this leads to a circularity that not only Kant would reject, but also loses all explanatory power. As Paul Formosa points out, "Kant thus wishes to know *why* our sociability turns unsocial. His answer is, of course, because of the propensity to evil rooted in the human species. Our unsocial sociability is the source of a strong incentive to adopt all sorts of *lower-order* evil maxims, but our radical evil is the embodiment of our choice of a supreme *dispositional* maxim. In other words, our radical evil *grounds*, but is not equivalent to, our unsocial sociability."[11] Consequently, the error in the attempt to equate unsocial sociability with human radical evil is that it integrates two distinct theses, and in doing so, inviting a vicious circle.

[11] Formosa, "Kant on the Radical Evil of Human Nature," 245.

2.3. Henry Allison's "deductive reading"

In *Religion*, Kant does not provide a formal argument for the thesis of radical evil, be it transcendental or not, despite the apparent need. Kant himself recognizes this need, when he says "[i]n order, then, to call a human being evil, it must be possible to infer *a priori* from a number of consciously evil actions, or even from a single one, an underlying evil maxim, and, from this, the presence in the subject of a common ground, itself a maxim, of all particular morally evil maxims" (*Religion*, 6:32). At the same time, he is "relieving himself the burden of providing it,"[12] suggesting that "[w]e can spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being, in view of the multitude of woeful examples that the experience of human *deeds* parades before us" (*Religion*, 6:33). This has been a deeply troubling matter for some readers, because the thesis of radical evil appears to be a synthetic *a priori* judgment—proving it would require transcendental deduction, according to the spirit of Kant's critical works. Even if, as others think, that the thesis of radical evil is not a synthetic *a priori* judgment, it is still evident that no amount of empirical data is sufficient to insist upon such universality and inextirpability. Consequently, some scholars of Kant have undertaken the task of somehow unearthing a formal argument that Kant merely hints at from the earlier sections of *Religion*.[13] One such attempt was made by

[12] For a detailed textual explanation on Kant's omission, see Seiriol Morgan, "The Missing Formal Proof of Humanity's Radical Evil in Kant's Religion."

[13] For other attempts at a reconstruction of a formal argument, see Paul Formosa "Kant on the Radical Evil of Human Nature," and Seiriol Morgan, "The Missing Formal Proof of Humanity's Radical Evil in Kant's Religion."

Henry Allison.[14] In this section I will discuss Allison's deduction, and the difficulties that it faces.

Allison's argument is based on Kant's illustration of the constitutive, or essential, features of human nature. By examining the human nature as described by Kant, Allison aimed to show that a good moral disposition is *conceptually* incompatible with human nature. In brief, his argument goes as follows:

(i) A good disposition, as opposed to an evil one, may be understood as the agent's supreme maxim that subordinates the incentives of self-love under those of the moral law.

(ii) For an agent with such a disposition, the "temptations" of self-love will be *automatically* dismissed in the presence of the requirements of morality, throughout the process of practical reasoning.

(iii) Such absence of temptation is impossible for humans as finite rational beings, for it is a necessary constituent of human nature to recognize the requirements of self-love and the ultimate desire for happiness in the process of practical reasoning.

(iv) Therefore, it is beyond the capacity of human beings to have a good disposition.

(v) Either a good or an evil disposition must be attributed to human beings, since we must somehow account for the enduring moral character of an agent, but not both (due to Kant's "rigorism").

(vi) Therefore, human beings cannot but have an evil disposition.

---

[14] See Henry Allison, "On the Very Idea of a Propensity to Evil," and *Kant's Theory of Freedom.*

2.4. Problems with the deductive reading

Though the deductive reading makes its best attempt at remaining loyal to Kant's original outlook by careful excavation of the text, it nonetheless suffers from difficulties of its own. One difficulty pertains to the soundness of the argument itself. There is little evidence to accept (ii) as true,[15] for the good disposition may instead indicate that even while incentives of self-love are ever present in demanding the agent's attention throughout the process of practical reasoning, the agent is capable, after deliberation, to silence the temptation and nonetheless choose to act upon respect for the moral law. As Markus Kohl adequately contrasts, "just as a propensity to evil does not eliminate our capacity to act from the recognition of duty, so a propensity to good does not eliminate our susceptibility to be tempted by considerations of happiness."[16] Although Allison's view attempts to solve the first problem by constructing a formal argument to prove the universality of human evil, the argument itself fails. Furthermore, it does not address the second and third problems at all.

An additional problem associated with Allison's view is that such an interpretation places the good disposition altogether beyond human reach. According to Allison, having a morally good disposition is impossible for humans on a conceptual level, because the concept of a good disposition and the concept of human nature is formulated as to be mutually exclusive. Consequently, it becomes metaphysically impossible for humans to have a good disposition, and the good

[15] See Formosa, "Kant on the Radical Evil of Human Nature," 241; and Markus Kohl, "Radical Evil as a Regulative Idea," 648-649.
[16] Markus Kohl, "Radical Evil as a Regulative Idea," 649.

disposition becomes "an unobtainable ideal of holiness."[17] This conceptual analysis of radical evil, however, trivializes the concept of radical evil.[18] To say that human beings are radically evil, in this account, is nothing more than to say that the nature of human beings are not the same as that of angels or saints. The most that can be achieved out of the analysis provided by Allison, is that according to Kant, no human as a sensuous being is immune to the requirements of self-love. Hence the analysis cannot by itself demonstrate *why* radical evil is a matter that ought to be taken seriously by all human beings.

3. The Regulative Reading

3.1. Motive and textual evidence

In this chapter, I endorse a possible interpretation of the doctrine of radical evil that focuses on the aspect of Kantian moral discipline, where Kant is understood as claiming not that it is an *actual, universal fact* that human beings are radically evil, but that every agent ought to *presuppose* that oneself has a propensity to evil, because that assumption is a cornerstone of each and every agent's lifelong journey of moral regeneration that Kant illustrates. This is a relatively novel approach to

[17] See Henry Allison, "On the Very Idea of a Propensity to Evil," 346-347.
[18] Allison refutes one "triviality objection" raised by Wood, but the objection raised here concerns a different aspect of triviality in Allison's deduction. For the triviality debate between Wood and Allison, see Wood, *Kant's Moral Religion*, 284; Allison, "On the Very Idea of a Propensity to Evil," 344-345; Wood, *Kant's Ethical Thought*, 287 and 402; and Allison, "Ethics, Evil, and Anthropology in Kant: Remarks on Allen Wood's *Kant's Ethical Thought*," 594-613.

the interpretation of radical evil, and I will first examine some textual evidence in favor of it. Then, I will discuss Markus Kohl's recent endeavor (2017) to sketch the general theory, and provide a formularized version of the main argument for the interpretation. Subsequently, I attempt to reinforce the regulative reading by explaining how the it successfully resolves the three interpretive problems, and also provide an additional reason to favor the regulative reading. Lastly, I will point out a potential difficulty that the regulative reading faces, and see if it can be resolved. Hopefully, the regulative reading as explained and supplemented here will provide a satisfactory explanation on how radical evil is better understood as a postulate for practical reason, because it not only deals with all three of the interpretative problems discussed earlier, but is also highly compatible with Kant's theory of moral discipline and self-perfection.

The striking force with which Kant stresses the universality and inescapability of human radical evil prompts one to consider the motive behind such forcible claims. Considering the meticulous and systematic nature of Kant's transcendental deduction and dialectics in his critical works, it is unjust to simply accuse Kant of "evading" or "omitting" to give a formal argument for the thesis of radical evil, and just as unfair to blame Kant for believing that empirical evidence was sufficient to prove his doctrine. Upon keener reading, we can find that Kant takes up a rather tentative attitude towards radical evil, and that Kant frequently uses the concept quite hypothetically. As Paul Formosa testifies; "[b]ut Kant himself, […] was fully aware of this, for he only states that '*if this is true*,' which 'everyone can decide by himself' (6:38), and not something along the lines of, 'as I have already proven.' Thus, it is not without reason that Kant often discusses humanity's radical evil only

in hypothetical terms."[19] But there are more than some insinuating hints in between sentences. Kant actually explicitly states that his doctrine of radical evil plays no significant role in surveying the principles of moral philosophy, in the following excerpt:

> The thesis of innate evil is of no use in moral *dogmatics*, for the precepts of the latter would include the very same duties, and retain the same force, whether there is in us an innate propensity to transgression or not. In moral *discipline*, however, the thesis means more, yet not more than this: We cannot start out in the ethical training of our conatural moral predisposition to the good with an innocence which is natural to us but must rather begin from the **presupposition** *(Voraussetzung)* of a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition; and, since the propensity to this [depravity] is inextirpable, with unremitting counteraction against it. [*Religion,* 6:50-51, emphasis added]

Here Kant degrades the importance of the actual truth of the thesis of radical evil, saying that the moral duties and their force would be the same whether there is, in fact, an innate propensity to evil in us or not. From this, we can see that Kant himself does not believe that he has actually proven the thesis of radical evil, because the factual truth of the thesis is irrelevant to his objective. Instead, Kant states that the thesis of evil serves as the cornerstone of moral discipline; that is, we must embark on the "ethical training" or "moral discipline" from the *presupposition* of radical evil. The reason for this presupposition is backed by Kant's claim to inscrutability: no finite rational being can fully and clearly discern one's own motives. Kant illustrates the path one must take for developing moral character, as follows:

---

[19] See Formosa, "Kant on the Radical Evil of Human Nature," 239.

> Assurance of this [*the new heart*] cannot of course be attained by the human being naturally, neither via immediate consciousness nor via the evidence of the life he has hitherto led, for the depths of his own heart (the subjective first ground of his maxims) are to him inscrutable. Yet he must be able to *hope* that, by the exertion of *his own* power, he will attain to the road that leads in that direction, as indicated to him by a fundamentally improved disposition. For he ought to become a good human being yet cannot be judged *morally* good except on the basis of what can be imputed to him as done by him. [*Religion*, 6:51]

From the two excerpts above, we may infer the motivation behind Kant's tentative use of the concept of radical evil. Rather than trying to demonstrate that all human beings have evil dispositions as an actual state of affairs, Kant could be dramatically accentuating that it is with such force that we should adopt a postulatory stance that every human being is by nature evil, because it is both the starting point of our progress for moral maturity and also the *sine qua non*, or essential condition, that should govern this progress throughout.

Here, I argue that there is another aspect of Kant's theory that creates a problem for the other interpretations, while adding to the plausibility of the regulative reading. Let us investigate Kant's claim to inscrutability a little further. According to Kant, for any agent, "the depths of his own heart (the subjective first ground of his maxims) are to him inscrutable." This is a crucial claim upon which the regulative reading stands. The Kantian reasoning behind this claim would be roughly that; the choices of incorporating an incentive into a maxim are the result of free will, which belongs to the faculty of the noumenal self. These choices, therefore, cannot be traced back from natural laws of causality or perceived through introspection. As phenomenal selves, we are not eligible to directly

perceive the process of our adoption of maxims, let alone the ultimate subjective grounds for them.

The inscrutability thesis is also congruent with our ordinary intuitions. How do we know that the psychology behind donations and charity is not something other than altruism, such as hypocrisy or the desire to be regarded superior to others? It is possible for us to mask our true intent with façades such as "the greater good" or "altruism," and even push self-justification to the point where we ourselves believe our own lies. Earlier in the first chapter I mentioned it is noteworthy that self-deception plays a part in the first two grades of the manifestation of evil, which are admittedly not difficult to find within ourselves. If we were to deny this and say we do not commit self-deception, how can we be sure that this isn't yet another self-deception to justify our own righteousness? The reason why we as moral agents are so prone to self-deception is perhaps because of our strong incentive of self-love; the regard for one's own well-being and happiness is so strong in us that it overrides our intention to see ourselves as who we really are. In other words, people prefer to think of themselves better than they actually are, due to the satisfaction and sense of well-being that a good self-conception offers. Kant exemplifies common instances of self-deception as follows:

> This is how so many human beings (conscientious in their own estimation) derive their peace of mind when, in the course of actions in which the law was not consulted or at least did not count the most, they just luckily slipped by the evil consequences; and [how they derive] even the fancy that they deserve not to feel guilty of such transgressions as they see others burdened with, without however inquiring whether the credit goes perhaps to good luck, or whether, on the attitude of mind they could well discover within themselves if they just wanted, they would not have practiced similar vices themselves,

> […] This dishonesty, by which we throw dust in our own eyes and which hinders the establishment in us of a genuine moral disposition, then extends itself also externally, to falsity or deception of others. And if this dishonesty is not to be called malice, it nonetheless deserves at least the name of unworthiness. [*Religion*, 6:38]

Kant illustrates how common it is for humans to "derive their peace of mind" through self-deception. This pervasive aspect of self-deception is a key element to the argument for the regulative reading, as will soon be discussed.

Still, one doesn't have to be an expert in the philosophy of mind to know that we are at least the best available authority on the states of our own minds. It is true that we may not fully know the true intent of ourselves, but it is often the case that we know our own intents better than those of others. If we are incapable of discerning the depths of our own heart, then to fathom the true intents of others would be even more unattainable. Hence, we can tweak the inscrutability thesis while preserving Kant's intuition, as follows: "no person can fully and clearly discern the motives and moral character of human beings, including oneself." I will call this the "common inscrutability thesis." In fact, Kant says basically the same thing when he describes the limitation of human judgment which can only trace or infer one's own disposition from assessment of observable maxims and actions (*Religion* 6:77), and claims that "no one can be certain how to estimate the character of any other man."[20] This declaration leads to yet *another* tension within Kant's theory of radical evil—Kant's common inscrutability thesis admits that no one can estimate the true moral character of oneself and others, while Kant himself seems to be claiming that every human being has a morally evil disposition. According to the

[20] Daniel O'Connor, "Good and Evil Disposition," 301.

common inscrutability thesis, Kant is not entitled to know the true moral character of himself and others, which means his claim that every human being has a morally evil disposition is unwarranted. There is no way that the anthropological reading or the deductive reading can explain the reason behind this tension, unless we accept that Kant was not, after all, claiming the factual truth of his judgment.

Returning to Kant's postulatory stance regarding the thesis of radical evil, consider the way Kant explains free will, Deity, or the immortality of the soul. In *Critique of Practical Reason* and *Fundamental Principles of the Metaphysics of Morals,* Kant explains how such concepts are required for the human morality to be meaningful as a whole. For instance, see the following excerpt from the second *Critique*:

> "Complete conformity of the will with the moral law is, however, *holiness,* a perfection of which no rational being of the sensible world is capable at any moments of his existence. Since it is nevertheless required as practically necessary, it can only be found in an *endless progress* toward that complete conformity, and in accordance with principles of pure practical reason it is necessary to assume such a practical progress as the real object of our will. This endless progress is, however, possible only on the presupposition of the *existence* and personality of the same rational being continuing *endlessly* (which is called the immortality of the soul). Hence the highest good is practically possible only on the **presupposition** (*Voraussetzung*) of the immortality of the soul, so that this, as inseparably connected with the moral law, is a **postulate** of pure practical reason." [*CPrR,* 5:122, emphasis added]

Kant's description of *the endless progress* and the required postulate of the immortality of the soul as practically necessary bears a striking resemblance to the way he describes moral regeneration and the role of the presupposition of human evil. Just as the endless progress for the complete conformity of the will with the

moral law is only possible on the presupposition of the immortality of the soul, the endless quest for good moral disposition, which is the agent's autonomous choice to make it the ultimate ground for the choice all maxims that respect for the law takes priority over self-love, is only possible through the presupposition of radical evil in the human heart. In both *Critique of Practical Reason* and *Religion Within the Boundaries of Mere Reason*, Kant uses the same vocabulary to describe the presupposition (*Voraussetzung*) of "the immortality of the soul" and "a depravity of our power of choice in adopting maxims contrary to the original ethical predisposition." *Voraussetzung* is often translated as "assumption, prerequisite, or requirement." In other words, the regulative reading regards human evil to be seen as a prerequisite, or "necessary hypothesis," that is required for the faculty of practical reason. The radical evil of humans, just like the existence of God and the ultimate triumph of good over evil, is a belief that we can neither prove nor disprove, but is required and justified for moral purposes via practical reason. For an agent who takes the task of moral discipline seriously and is committed to a lifelong prioritization of the respect for the moral law, it is necessary to live *as if* the thesis of radical evil were true, because of its immense importance in our practical and ethical lives. The reason for the necessity of such postulate is that the human being is so prone to self-deception that even a slight hint of self-righteousness may lead to self-conceitedness, which is one of the least desirable consequence for moral discipline.[21]

[21] As might be expected, some may protest that the textual evidence provided above do not suffice to regard radical evil as a presupposition. I do not deny that there are other passages throughout Kant's text in which he seems to make claims that support the other interpretations discussed in the former sections. It is equally possible, if one were to peruse the many literature that Kant authored, to find

Recall that in chapter 1 we examined Kant's idea of moral self-perfection, and the two-stage model of moral reform that ultimately aims to attain it. Being free rational agents, humans are given the duty of moral self-perfection (MM 6:386–87), which is to engage in this never-ending endeavor, regardless of one's individual moral experiences and adoptions of good or evil maxims. Even if one succeeds in once adopting and acting from a good maxim, one should continuously build the inner strength to resist the constant temptation. Indeed, this illustration of the task of the "moral ascetic" bears a distinctly "Sisyphean" impression, as Allison observed, in the sense that this task is a never-ending process of withholding the weight of the incentive of self-love (although its result may not be as fruitless as the punishment of Sisyphus). What makes this task even more burdensome is the inscrutability thesis, according to which no finite rational agent is entitled to fully and clearly perceive one's true motives and moral character. All we are entitled to "know" is that we are either morally good or morally bad, for the rigorism thesis excludes moral middle and states there is no gray zone in moral character. As for our current moral status, we can only assess ourselves by tracing and inferring from the "upper hand" that we gain over time, as we continue to take on the second stage of moral reformation. In other words, if we make constant conscious attempts to incorporate only the incentive of the respect for the moral law into the choice of our maxims and gradually it becomes more often the case that our actions cohere with the law, there is a possibility that perhaps we are on

---

excerpts that appear to reject each of the three interpretations. Under such circumstance, the three problems of interpretation mentioned earlier serve as canons to evaluate which interpretation best accounts for the matter at hand. Luckily for the proponents of the regulative reading of radical evil, textual evidence is not the sole element that gives reason to accept that radical evil has the status of a postulate.

the right path and our moral progress is indeed a manifestation of our moral conversion to recover the supreme maxim that prioritizes the moral law; but we can never be certain of it.

I have briefly discussed earlier that Kant places more emphasis on the role that radical evil plays in moral discipline than in moral dogmatics, and that Kant's theory of radical evil and the process of moral regeneration should not be understood apart from each other. The problem that lies within the motivational level of the anthropological reading is that it underestimates the weight of Kant's commitment to moral regeneration. Allison, on the other hand, recognizes this weight with remarkable insight: He recommends that "the concept of a propensity to evil also plays a crucial role in Kant's account of moral development," because "the doctrine of radical evil, in the form of an original propensity to evil, not only defines our moral condition but also sets the moral agenda for imperfect beings such as us."[22] That being said, the regulative reading can be associated with Kant's commitment to moral regeneration more than any other interpretation. Consider the relationship between moral conversion and the postulate of radical evil. In the two-stage model of Kant's moral reformation, the first stage consists of a "single and unalterable decision" of an agent to restore the reversed subordination of incentives in his supreme maxim. But in order to make a resolution to "restore" something, one must first admit that the status quo is in reverse and needs to be fixed. According to the regulative reading, the postulate of radical evil is virtually the same as *accepting* that the agent's current moral disposition is evil. This

[22] Allison, "On the Very Idea of a Propensity to Evil," 346.

presupposition, therefore, plays a pivotal role in the sense that it is a prerequisite for the first stage of moral reformation. Although the postulate that oneself has a morally evil disposition does not ensure that one has indeed also made the fundamental decision to restore the lawful order of incentives, it certainly does serve to mark a possibility.

3.2. Presupposition and belief

At this point I would like to attend to an issue that needs to be resolved in order for the regulative reading to stand, regarding whether there is a meaningful difference between the two propositional attitudes of *believing P* and *presupposing P*. Given that presupposing *P* requires that the agent live *as if P* is true, what epistemic quality differentiates presupposition from belief? If there cannot be drawn a significant distinction, the presupposition of radical evil could turn out to be none other than belief of it. Since believing something is generally understood as taking it to be objectively true, such reveal would incapacitate the regulative reading's claim that Kant was not trying to assert the objective truth of the thesis of radical evil. Although the matter at hand has been the topic of extensive debate in both areas of epistemology and moral psychology, for the purposes of this thesis I will limit my discussion to Michael Bratman's distinction between belief and

acceptance (1992)[23], and see whether the Kantian conception of belief and presupposition can also be distinguished in a similar manner.

Bratman, in analyzing the different cognitive attitudes that guide our practical reasoning and action, claims that belief and acceptance are two distinct attitudes. He suggests the following qualities that differentiate acceptance from belief. First, while reasonable belief normally "aims at truth," meaning that it is usually formed with some evidence that stands for the truth of what is believed, reasonable acceptance can be governed by practical considerations that need not necessarily involve commitment to the actual truth of what is accepted. Second, while belief is not the subject of our direct voluntary control, acceptance is; one can accept some premise instantly following one's voluntary decision to do so. Third, while what one believes is context-independent (i.e. one cannot, at the same time, believe that *P* in one context and not believe that *P* in another without loss of consistency in one's set of beliefs), what one accepts can vary across different contexts without threatening the soundness of one's practical reasoning skills. In short, for an agent to *accept* that *P* means that the agent voluntarily takes *P* as a premise in one's practical reasoning, by regarding *P* as true in a given context, not necessarily with evidence for its truth but in response to pragmatic considerations, or "practical pressures" that render the acceptance of *P* to be of practical interest to the agent. Bratman exemplifies these features of acceptance through various cases and enumerates the pragmatic reasons that one may have for accepting a premise. One of them involves the "asymmetries in the cost of errors," in which an agent accepts

[23] For other works regarding the distinction between belief and acceptance, see van Fraassen (1980), Stalnaker (1984), Bratman (1992), Cohen (1992).

a premise as a result of a risk assessment in one context but not another, because of the different degrees of risk that one's practical reasoning carries. See the following example:

> I am planning for a major construction project to begin next month. I need to decide now whether to do the entire project at once or instead to break the project into two parts, to be executed separately. The rationale for the second strategy is that I am unsure whether I presently have the financial resources to do the whole thing at once. I know that in the case of each sub-contractor—carpenter, plumber, and so on—it is only possible at present to get an estimate of the range of potential costs. In the face of this uncertainty I proceed in a cautious way: In the case of each sub-contractor I take it for granted that the total costs will be at the top of the estimated range. On the basis of these assumptions I determine whether I have at present enough money to do the whole project at once. In contrast, if you offered me a bet on the actual total cost of the project—the winner being the person whose guess is closest to the actual total—I would reason differently. [24]

Within circumstances such as the above where the stakes for my judgment are high, I could take extra precaution in my reasoning and accept premises that will act as safety measures for my reasoning and actions in some way, even though I do not take those premises to be actually true. We may easily conceive a similar scenario for the Kantian presupposition of radical evil. In Kantian moral discipline, the cost of error for the judgment of one's moral character is extremely high. If one were to judge (wrongly) that oneself has a morally good disposition, one is prone to self-deception which is crucially detrimental to moral growth. Therefore, the moral agent could take extra precaution in her practical reasoning and voluntarily accept the thesis of radical evil as a premise in her practical conduct and moral regeneration, by regarding it as true in (and only in) the context of her own moral

[24] Bratman, "Practical Reasoning and Acceptance in a Context," 6.

reform, not necessarily with evidence for its truth but because the acceptance of radical evil is of moral interest to the agent.

I have mentioned that the agent may regard the thesis of radical evil as true *only* in the context of her own moral reform, because understanding acceptance as a context-relative attitude also helps explain how we can presuppose the thesis of radical evil but at the same time engage in interpersonal activities and emotional commitments to others. Given that we believe that each and every human individual in this world is thoroughly and irrevocably evil at heart, how could we stop ourselves from becoming misanthropes? How would we be able to create emotional bonds with our friends and family, and foster sincere relationships with others, if we harbor the belief that those that we love and care for are in fact downright egocentric and self-serving? In such cases, either we would maintain our social lives and become somewhat hypocrites in the sense that our actions and behaviors toward people do not reflect what we actually believe them to be, or become recluses and refrain from interpersonal relationships. Neither case seems hardly compatible with the life of a virtuous person, with sincerity, integrity, benevolence, tolerance, and many such virtues that make a person's character admirable. How, then, could the thesis of radical evil be said to contribute to our becoming morally better persons? This is arguably one of the reasons why Kant's thesis of radical evil seemed so absurd to many people; the belief that all humans are evil seems to be the type of conspiracy theory that puts ordinary life at risk. However, this absurdity ensues because belief is a context-independent attitude. *Accepting* the thesis of radical evil, on the other hand, is open to context relativity and requires only that the agent regard it to be true at a certain context: in this case,

the context of one's moral discipline. The premise of radical evil then serves only as a safety measure to keep oneself from yielding to complacency and self-deceit. Outside of said context, the thesis of radical evil need not exercise any influence over one's pragmatic reasoning and actions, so the agent could wholeheartedly commit to interpersonal relationships without taking into account the presupposition that the people they encounter are evil in character.

But all of this applies only when Kant's presupposition (*Voraussetzung*) can indeed be considered equal, or at least sufficiently similar, to reasonable acceptance. Can Kantian presupposition be identified with acceptance? Bratman certainly seems to think so, since in exemplifying instances of reasonable acceptance he introduces the Kantian example of the presupposition of free will, as follows:

> Having reflected on issues about free will I am perplexed about whether I have it. Yet 1 still must on occasion deliberate about what to do. When I do I need to accept that what I will do is to some extent up to me. *I need to accept that I have a kind of free will I do not believe I have.* [emphasis added] And it is hard to see how such acceptance could fail to be practically rational; for its absence would preclude any practical reasoning at all. [25]

Here, my acceptance (or, in Kantian terms, presupposition) of free will is based on the pragmatic consideration that without such acceptance, it will be difficult for me to carry out my everyday conducts with all my other practical reasonings. Therefore, even without the justified belief that I indeed have free will, I accept it as a postulate of practical reason and proceed with my life. This presupposition of

[25] Ibid., 8.

free will is influenced by another variety of practical pressure unlike that which affects the presupposition of radical evil, but nonetheless both attitudes fall into the same category of acceptance. However, it remains to be seen whether such distinction can indeed be made within Kant's practical theory with textual support. To examine Kant's conception of various propositional attitudes, I refer to Andrew Chignell's research (2007) on the Kantian system of belief.

According to Chignell, belief in Kant can be understood either in a broad sense as a systematized set of propositional attitudes related to what is roughly described as holding something for true, which is more aptly translated as "assent" (*Fürwahrhalten*)—as the genus of which most other positive propositional attitudes are the species—or it can be understood in a narrow sense as "Belief" (*Glaube*) by which Kant denotes a particular species of assent. I will briefly explain how the overall system of assent is construed, and then show how a parallel can be drawn between belief/acceptance (as distinguished by Bratman) and its Kantian counterparts. Kant states in "the Canon of Pure Reason" in the first *Critique* that the assent of a proposition requires being "sufficiently (*zureichend*) grounded," which is similar to the modern term "justification" of a belief or judgment. This epistemic sufficiency can be either evaluated of *objective* grounds (*Gründe*) or *subjective* grounds (*Ursachen*), and the species of assent can be categorized by which type of ground and/or the other is sufficiently assented of a given proposition. Regarding what makes these grounds sufficiently assented, Chignell lists five features of objective sufficiency and introduces two different kinds of subjective sufficiency. In short, what makes an objective ground sufficient are those aspects that are usually seen to justify a judgment to be objectively true, such

as having a probability to a certain degree of being true, or being "intersubjectively valid" so that any rational agent in the same epistemic condition is likely to assent to its truth, or causing involuntary cognitive attitudes so that any rational agent who acquires sufficient objective grounds for *P* would typically find their assent for *P* to follow. What can serve as objective grounds is mostly what we would normally call empirical evidence, such as "perceptual, memorial, and introspective states, as well as other sufficient assents we already hold (the results of inductive and deductive arguments, assents about what others have testified, assents about one's experiences, and so forth)." Notice how Bratman's specification of belief coheres with Kant's objectively sufficient assent. Both attitudes are formed in response to some evidence that stands for the truth of what is believed or assented, both have the tendency to retain its probability regardless of variation in context, and both accompany involuntary responses of belief-formation from epistemic agents. At this point, it seems safe to assume that belief (as specified by Bratman) and objectively sufficient assent are qualitatively similar to a large extent.[26]

Meanwhile, the *subjective* sufficiency of an assent is again classified into two types; the first type of subjective sufficiency is the requirement of "the subject's own determination that the assent is based on sufficient objective grounds," that the epistemic agent would, upon reflection of her reasoning, cite the given ground as the sufficiently objective ground for her assent. This can easily be understood as a constraint that prevents mere epistemic coincidence from counting as reasoned

[26] Species of assent can be further classified by whether the grounds for the assents are objectively sufficient or *insufficient*, and again, in each case whether they are subjectively sufficient or *insufficient*. For more detailed accounts of the four kinds of assent, see Andrew Chignell (2007).

assent. What we are more interested in is the second type of subjective sufficiency. This type of sufficiency draws directly from the practical "interest" of the agent, and "needs of reason" that make certain assents desirable for agents. Such interest is *nonepistemic* in the sense that it makes an assent desirable for an agent not by its epistemic values such as objective truth, but by pragmatic considerations. Chignell explains that subjective sufficiency in this sense makes certain assents "rationally acceptable for certain people in certain contexts," and classifies assents that are only subjectively sufficient in this second sense as Belief (*Glaube*). As can be expected, Kant's presupposition of human free will, Deity, and immortality of the soul fall into this category.[27] Chignell observes that Kant's notion of Belief (*Glaube*) is akin to the modern conception of "acceptance," and hence bears a much narrower meaning than the English word belief. These characteristics of Belief (*Glaube*) parallel those of acceptance as specified by Bratman. Both attitudes are formed in response to pragmatic considerations as opposed to epistemic responses, both are context-relative, and are objects of direct voluntary control. Therefore, I contend that we can safely identify Kantian Belief (*Glaube*) with acceptance; there is now a clear distinction between propositional attitudes of presupposition (*Voraussetzung*) and belief. Assuming that the presupposition of radical evil is a voluntary, context-relative acceptance of the thesis of radical evil into the agent's practical reasoning within the context of moral discipline influenced by its pragmatic concerns, I will move on to the next section and discuss Markus Kohl's argument for the regulative reading.

[27] More specifically, they are classified as Moral Beliefs, and Chignell, taking the example of human equality, also endorses that such Moral Beliefs can be approved from a modern perspective.

3.3. Markus Kohl's moral ascetic argument

With a similar rationale, Markus Kohl recommends that we understand Kant's radical evil as a "regulative" concept. He derives the idea from Kant's regulative concepts of pure reason, and suggests that Kant's radical evil is a regulative concept of practical reason. Kant, in the *Critique of Pure Reason,* introduces the distinction between regulative concepts and constitutive concepts. The constitutive concepts or principles belong to the faculty of understanding, to which all experience must conform to. On the other hand, the concepts proper to the faculty of reason are *regulative* in the sense that they "present us, not with objects corresponding to them, but rather with a task: the never ending progress of empirical enquiry whose ideal terminus […] can only be approached asymptotically."[28] Kohl claims that Kant's intention, when proposing the doctrine of radical evil, was not to claim that the universality of human radical evil is an objective truth, but rather to claim that we ought to assume human beings as evil, as a regulative principle. According to Kohl's understanding, "[r]egulative principles concern a proposition that we can neither prove nor disprove, that is, that might be true or false for all we can know with certainty."[29]

Inspired by the first-person perspective of the agent who is faced with the task of moral discipline, Kohl describes the epistemic conditions of such moral agents and demonstrates how the postulate of radical evil is most suitable to the agent

[28] Michael Friedman, "Regulative and Constitutive," 73.
[29] Markus Kohl, "Radical Evil as a Regulative Idea," 656.

because it minimizes the various risks that impair moral discipline. Here I provide a structuralized 'moral ascetic' argument suggested by Kohl, as follows:

(1) For the agent who takes up the task of moral discipline seriously and is earnestly committed to making oneself a better person throughout one's life, one of the following two is true (given the rigorism thesis): (i) one currently has a morally good disposition; (ii) one currently has a morally evil disposition.

(2) No finite rational agent can fully and clearly discern one's own motives and moral character.

(*The Inscrutability Thesis*)

(3) Since the agent cannot objectively perceive the state of one's moral character [due to (2)], one can choose to presuppose one of the following three attitudes: (i) that one currently has a morally good disposition; (ii) that one currently has a morally evil disposition; or (iii) suspend the judgment of one's own moral character.

(4) If the agent chooses to presuppose (i) or (iii), there is a probability that the agent would risk oneself to self-deception.

(5) The agent who takes up the task of moral discipline seriously and is earnestly committed to making oneself a better person throughout one's life would have a reason to minimize the probability that one succumbs to self-deception.

(6) Therefore, it is most reasonable (out of moral interest) for the agent to presuppose (ii).

The basic argument takes the form of disjunctive syllogism. Under the assumption that the agent has accepted to take on the task of the moral improvement, the argument provides that the best mindset for the agent is to presuppose that oneself has yet a morally evil disposition, for it is the only way to take the best precautions to minimize the risk of falling into the trap of self-deception that gravely impairs moral improvement. We have seen in the first section how the two grades of the manifestation of evil involve self-deception. The

pervasive and elusive nature of self-deception within the manifestation of evil makes it critically detrimental to moral discipline.

3.4. Evil as self-deception

There are different forms of self-deception that attitudes (i) and (iii) may each invite. First, choosing (i) as the mindset for moral development may invite complacency as a form of self-deception. If an agent is to believe that she currently has a morally good disposition, the only improvement she can make in her moral character is to foster and stabilize her strength of will in implementing actions that comply with the law. However, as has been mentioned earlier, the mechanical habituation of the execution of actions is thoroughly compatible with an evil disposition; the agent cannot eradicate the possibility that she is involved in a deep self-deception that leads her to believe that she is acting entirely out of the supreme maxim that prioritizes respect for the moral law, while her actions and decisions are in fact feeding her self-satisfaction of looking all good and saintly.

Adopting (iii) as the initial mindset does not help either, because such agnostic stance invites a type of "moral laziness" as a form of self-deception. Suspending judgment on one's moral character may easily lead to procrastinating the need for moral development, because the moral agent who chooses (iii) can at best assume that she might have an evil disposition, and equally that she might have a good disposition. This attitude of acknowledging a mere possibility significantly reduces

the urgency to take immediate action and strive to improve. An analogical example would be a comparison between the two beliefs I may have the night before an important exam. I may believe that I *might* fail the exam if I go to bed early because I haven't studied enough, or I may believe that I *will* indeed fail the exam if I go to bed early. Which attitude should I choose to ensure the best possible score on the exam? The former belief leaves the back door open by allowing me to believe that I might equally pass the exam due to luck or some intellectual superiority I might have over other students.

That said, the weight of the subjective commitment that the moral ascetic places on moral development is incompatible with an attitude that diminishes the chance to attain moral character and increases the chance for self-deception and complacency. In other words, neither presuming that one has a good moral disposition nor suspending that judgment has any benefit whatsoever for the moral interest of the ascetic, while presuming the opposite promotes the best chance for the moral interest of the ascetic. Kohl states that in presupposing an evil disposition in oneself, a moral ascetic has "*nothing to lose but everything to win.* She has nothing to lose because she must engage in incessant counteraction anyway, regardless of whether she acts from a good or evil maxim. She has everything to win because the presupposition that she has an evil character facilitates *the most effective, sincere, and whole-hearted way* of ensuring that one reliably acts for moral reasons and of counteracting pervasive threats such as self-deception and moral complacency."[30] The assumption, or belief, that one is disposed to prioritize

[30] Kohl, "Radical Evil as a Regulative Idea," 666, emphasis added.

the incentive of self-love over the incentive of respect for the moral law acts as not only a motivational starting point but also a safeguard, a buoy for those in the process of the voyage of moral development to avoid invisible reefs or hazards hidden beneath the observable surface. Undeniably, a major motivation of the regulative reading is that the more serious an agent is involved in the process of moral development, the more one realizes the gravity of the threat of moral self-deception. This is why Kant insists that "according to the cognition we have of the human being through experience, he cannot be judged otherwise, in other words, we *may presuppose* evil as subjectively necessary in every human being, *even the best*." (*Religion*, 6:32, emphasis added) We can understand Kant as suggesting that the best way to prevent the threats of self-deception is to commit oneself to an unyielding renunciation of the possibility that oneself has a morally good disposition, contrary to what may actually be the case. On this account, radical evil serves to protect the moral ascetic through a *freely chosen* blindfold to foster modesty and foil arrogance.

3.5. Merits and practical implications

In the earlier two sections, I have explained Markus Kohl's attempt to view the role of radical evil as a necessary hypothesis in Kant's theory of moral development. Although Kohl's argument plays a leading role in the reasoning of the regulative reading, since it is an original attempt, it falls short of supplementary

elements to strengthen the interpretation such as its theoretical merits or possible objections to the argument itself. Therefore, in the following sections I aim to further contribute to the regulative reading by presenting the theoretical benefits of the regulative reading, and an additional reason to favor the it.

To begin with, as I mentioned earlier, the fourth tension regarding the common inscrutability thesis which would be problematic for the anthropological and the deductive readings does not pose any threat to the regulative reading. Moreover, none of the three interpretative problems arise. The problem of universality is immediately eliminated because the regulative reading does not regard Kant as asserting the universal truth of the thesis of radical evil, and thus requires no actual proof, but only the presupposition of it.

Secondly, the seeming conflict between human evil being a natural property despite its imputability to humans is not a problem for the regulative reading, because it is part of its presupposition. See the following excerpt:

> "He is evil *by nature"* simply means that being evil applies to him considered in his species; not that this quality may be inferred from the concept of his species ([i.e.] from the concept of a human being in general, for then the quality would be necessary), but rather that, according to the cognition we have of the human being through experience, he cannot be judged otherwise, in other words, **we may presuppose evil as subjectively necessary in every human being, even the best.** [*Religion*, 6:32, emphasis added]

In light of the regulative reading, we may interpret the above passage to mean the following. Due to the common inscrutability thesis, we cannot discern the true moral disposition of any agent; but because of our high susceptibility to self-deception, it is required that we presuppose that ourselves have a morally evil

disposition, to take on the duty of moral self-perfection. Since this applies to every human being, we may presuppose that *every* human being has an evil supreme maxim, and in the sense that this is applicable to all human beings, we may call radical evil to be applicable to our species, and hence, our "nature." Seen this way, the presupposition that evil is entwined with human nature does not reduce our responsibility for evil in moral discipline. For an agent to presuppose the thesis of radical evil is to volunteer to take full responsibility for one's exercise of the power of choice and the possibility of self-deception and the evil that follows from it. Our presupposition that we are innately inclined to prioritize self-love does not justify that we can loiter in our weaknesses; rather, this presupposition was taken precisely because we want to be the opposite, to extricate ourselves from this weakness through arduous training.

The third problem regarding the inextirpability of human evil and our duty (and hence, ability) to become morally better persons can also be explained by the relationship between the inscrutability thesis and our duty of moral self-perfection. As seen earlier, our duty of moral self-perfection requires us to engage in the incessant task of adopting and acting from good maxims, and ultimately to change our fundamental cast of mind. But due to the inscrutability thesis, no moral agent, even one who has achieved a considerable degree of moral maturity, can be certain of one's own moral disposition. To avoid the complacency that one has already reached a state of moral perfection, it is required for the agent to postulate that the propensity to prioritize self-love over moral law is ever present in oneself, which makes the inextirpability of human evil another necessary presupposition that

follows from the mechanism of moral discipline. Kant illustrates this point in the following excerpt:

> For him who penetrates to the intelligible ground of the heart (the ground of all the maxims of the power of choice), for him to whom this endless progress is a unity, i.e. for God, this is the same as actually being a good human being (pleasing to him); and to this extent the change can be considered a revolution. For the judgment of human beings, however, who can assess themselves and the strength of their maxims only by the upper hand they gain over the senses in time, the change is to be regarded only as an ever-continuing striving for the better, hence as a gradual reformation of the propensity to evil, of the perverted attitude of mind. [*Religion,* 6:48]

An omniscient being such as God would be able to discern the true disposition of humans; however, due to the inscrutability thesis, the judgment of human beings as phenomenal selves is limited in the sense that one can never be absolutely certain of one's own moral improvement, even if all the visible evidence stands in favor of it. To minimize the risk of complacency and a false self-conception, one needs to presuppose that oneself is always still working towards the good, and never has already attained it, and in this sense radical evil is presupposed to be inextirpable.

Now I suggest an additional reason to favor the regulative interpretation over others. I mentioned earlier that it would be a further advantage if an interpretation of radical evil can also provide some insights into the practical implication of such a conception of evil; what Kant's motive was in depicting human evil in such a radical way, and why it is meaningful to understand human evil as Kant proposes. The regulative reading can do just this, because it is the only account that makes sense and sketches a compelling image for anyone who wishes to become a

morally better person, even outside the Kantian framework and without direct commitments to Kant's specific ethical conceptions. Meanwhile, other interpretations in the literature make no sense outside Kantian exegetics—there is little reason to accept Kant's teleological view of human history or Kant's definition of human nature, without which neither the anthropological nor the deductive reading can stand.

On the other hand, even on a commonsensical level, the human moral condition as depicted by the regulative reading is still reasonable, because most of the precepts of Kant's theory of radical evil and moral discipline follow from the inscrutability thesis: the human limitation that we cannot fully and clearly perceive our own intentions and that we are prone to self-deception. Arrogance, complacency, and blindness to one's own faults are traits that even those who are not committed to a Kantian moral discipline dislike to be associated with and yet are quite easily tempted by. Therefore, the regulative reading provides insight into what we must heed to as long as we want to be "better persons." Just as much as we try to cultivate other virtues such as kindness, courage, and wisdom, we should pay equal attention to our self-conception which may easily turn complacent during our endeavor to become better persons.

In his book *The Road to Character*, dedicated to address the importance of humility in moral improvement, David Brooks depicts a similarly motivated view of the human moral condition. While underscoring the gravity (and scarcity) of humility, Brooks illustrates a cultural and intellectual tradition that he believes is

now all but forsaken, called the "crooked timber" tradition[31], which emphasized the universal weakness in humans. However, the acknowledgment of one's weakness did not weaken humans to succumb to despair for their limitations. On the contrary, it was a tradition that "held that each of us has the power to confront our own weaknesses, tackle our own sins, and in the course of this confrontation with ourselves we build character."[32] Brooks deplores the contemporary zeitgeist which places too much emphasis on a grandiose self-conception, and warns that this pervasive self-centeredness leads to selfishness and unwarranted pride over others. He then gives an interesting portrayal of the human morality that bears a striking resemblance to Kant's outlook, as follows:

> Some perversity in our nature leads us to put lower loves above higher ones. […] We all know the love you have for the truth should be higher than the love you have for popularity. […] But we often put our loves out of order. If someone tells you something in confidence and then you blab it as good gossip at a dinner party, you are putting your love of popularity above your love of friendship. If you talk more at a meeting than you listen, you may be putting your ardor to outshine above learning and companionship. We do this all the time. [33]

Brooks' intuition corresponds with Kant's description of reversed incentives, where a love for oneself is prioritized over what rightfully ought to be placed greater weight. In order to set this priority of love straight, Brooks commends the humble self-conception of the "crooked timber" tradition, and encourages a lifelong combat and inner struggle against selfishness. The rest of his book is

---

[31] Incidentally, this nickname given by Brooks originates from a quote by Kant, "out of the crooked timber of humanity, no straight thing was ever made," from *Idea for a General History with a Cosmopolitan Purpose* (1784), Proposition 6.
[32] Brooks, *The Road to Character*, xiv.
[33] Brooks, *The Road to Character*, 11.

packed with biographies of actual people who have lived accordingly, which would serve well as empirical evidence that such perspective on moral development is neither an absurd nor an impractical idea. In the following section, I further probe the intuitive disagreements over moral development through examining a possible objection to the regulative reading.

3.6. The cancer argument

I close my thesis by addressing an objection that may be raised to weaken the practical plausibility of the regulative reading that I have just proposed. The objection is that the force of the moral ascetic argument may not be as strong as the proponents of the regulative reading hope to be. Consider the following argument:

(i) We ought to obtain *A.*

(ii) In order to obtain *A*, we must presuppose *B*.

(iii) Therefore, we must presuppose *B*.

Then, consider the next adaptation:

(i') We ought to obtain moral self-perfection through constant self-training.

(ii') In order to obtain moral self-perfection through constant training, we must presuppose that we are radically evil.

(iii') Therefore, we must presuppose that we are radically evil.

And finally, the following:

(i") We ought to obtain physical health through constant self-care.

(ii") In order to obtain physical health through constant self-care, we must presuppose that we have terminal cancer.

(iii") Therefore, we must presuppose that we have terminal cancer.

Intuitively, the last argument is far from convincing. In order for the above arguments to gain a sufficient degree of persuasiveness, it is important that the postulate suggested in the second premise is forceful enough to be considered a contributing factor to the acquisition of *A*. However, there are other possibilities to obtain physical health via constant self-care without the drastic presupposition suggested above, and therefore the argument fails. It may be the case that the presupposition of radical evil is the same; why should we postulate that we are morally corrupt to the core in order to ensure that we avoid self-deception and obtain moral self-perfection, if especially the moral self-perfection as pictured by Kant requires such strict adherence to the moral law?

First, it should be noted that this objection may weaken only the last point I made. Early on, I suggested that it would be a further advantage if an interpretation of radical evil can also provide some insights into the practical implication of such a conception of evil; why it is meaningful to understand human evil as Kant proposes. The above objection does not jeopardize the entire regulative reading, but only the tenability of its practical implications. Admittedly, this is a problem that arises for those who wish to gain practical insight from a Kantian perspective; we know that for Kant himself, such moral perfectionism is not an absurd idea at all. But for those who, as I illustrated in the earlier section, wish to gain insight from Kantian radical evil in their attempts to become morally better persons, such a

drastic presupposition may not be so appealing. Shouldn't there at least be some kind of a psychological reward, an acknowledgment that oneself is indeed becoming a better person, to keep fueling the motivation to become a better person? Must we suppress our self-esteem to such extent? To this, I must admit that this is the point where intuitions diverge. Think about professional athletes, how they are all deeply committed to being good sportspeople but their methods for training vary person by person. To one, an austere mindset that strives for perfection without tolerating a single deviation from the hard routine may work best; to another, a more lenient, relaxed mindset that allows occasional recreation and leisure may work best. Likewise, among those who wish to become morally better persons, a strict mindset that does not tolerate even the slightest possibility for self-deception may work best for one person; to another, a more optimistic view of the self that sees oneself as gradually becoming a better person may work best. I do not mean to insist that the former method is the better of the two, but I wish to illustrate the point that the former mindset is equally reasonable as the latter.

Put another way, the comparison between the two mindsets is to compare which of the following two is the more discouraging: the despair that may come from presupposing that one is radically evil and unable to obtain moral perfection no matter how hard one tries, and the complacency and moral idleness that may come from presupposing that one is morally good, or at least becoming so. As we have seen so far, at least for Kant himself, the latter idea is far more discouraging than the former. If, on the contrary, the former idea is more discouraging than the latter, it would be unreasonable to accept the Kantian conception of evil. However, the idea that one may not ultimately obtain some ideal perfection may not be as

discouraging as it seems. In fact, it is not an uncommon endeavor to attempt something, knowing that its perfect acquisition is impossible. Consider the action of drawing a circle well. Anyone with elementary geometric knowledge knows that drawing a perfect circle is impossible in our reality; but it is not absurd to try to be able to draw the circle closest to perfect as possible. So, to set some ideal objective while knowing that it is unattainable due to my imperfection does not seem to be such an obstacle to my motivation or desire to keep trying to achieve it. Similarly, to set an ideal objective on moral perfection while knowing that my moral commitment is too weak to achieve it does not have to critically dishearten me from still wanting to be a morally good person. As British writer Henry Fairlie wrote; "If we acknowledge that our inclination to sin is part of our natures, and that we will never wholly eradicate it, there is at least something for us to do in our lives that will not in the end seem just futile and absurd."

An example of this type of motivation for moral discipline can be drawn from Joel Feinberg's critique against psychological egoism (1999). Feinberg takes the fictional case of Abraham Lincoln and the drowning pigs to draw an objection to psychological egoism, the claim that all human actions can be seen to be motivated by selfish desires. The anecdote goes as follows:

> Mr. Lincoln once remarked to a fellow-passenger on an old-time mud-coach that all men were prompted by selfishness in doing good. His fellow-passenger was antagonizing this position when they were passing over a corduroy bridge that spanned a slough. As they crossed this bridge they espied an old razor-backed sow on the bank making a terrible noise because her pigs had got into the slough and were in danger of drowning. As the old coach began to climb the hill, Mr. Lincoln called out, "Driver, can't you stop just a moment?" Then Mr. Lincoln jumped out, ran back and lifted the little pigs out of the mud

> and water and placed them on the bank. When he returned, his companion remarked: "Now Abe, where does selfishness come in on this little episode?" "Why, bless your soul Ed, that was the very essence of selfishness. I should have had no peace of mind all day had I gone on and left that suffering old sow worrying over those pigs. I did it to get a peace of mind, don't you see?" [34]

Feinberg uses this story to claim that contrary to Lincoln's assertion, there must have been an altruistic motive in Lincoln's aid to the pigs. Whether it is Lincoln's psychological egoism or Feinberg's objection that is actually true of humans is not a concern for the regulative reading, but we are interested in Lincoln's claims and his intentions behind them. Why would Lincoln hold the sweeping claim that all humans are thoroughly selfish, and when he has done a generous deed, why would he expose himself to have acted out of a selfish motivation, when such a motivation would have stayed concealed had he not said so, and he could have been praised for his charitable character? Is it because Lincoln was so misanthropic and unsociable that he abhorred all human beings and their conducts? Although this anecdote is possibly fictional, it is based on a real person of whom we have plenty of biographical information. The general reputation of Abraham Lincoln is that he was someone of noble and respectable character, contributing to the freedom of slavery and the establishment of constitutional rights.

There is a message that we can infer from the dissonance between Lincoln's claims and the fact that this story is based on Abraham Lincoln. It is that perhaps Lincoln harbored a belief that is similar to the postulate of radical evil and had a deep commitment to becoming a morally better person; therefore, in order to avoid

[34] Feinberg, "Psychological egoism," 497.

the reputation that he does not deserve and would only incur self-complacency, he made claims about himself that coheres with the presumption that he is in fact a selfish man, contrary to what he may seem like. These claims were not made because Lincoln was a man of self-hatred and a morbid self-esteem, but in fact the very opposite: because he was a person who respected moral good and strived to be a better person. Indeed, this may not be the stance that everyone is fit to adopt. I do not wish to argue that the presupposition of radical evil is the best possible method for moral growth, but only that there is enough reason to consider that Kantian radical evil, regarded as a postulate for practical reason, makes meaningful contributions to moral development.

Conclusion

The contributive aim of this thesis is threefold. First, I have identified the three major puzzles that arise regarding Kant's doctrine of radical evil. These puzzles serve as a standard by which all attempts at interpreting radical evil can be evaluated in comparison. Second, based on this criterion I assessed two earlier interpretations of Kant's radical evil by Allen Wood and Henry Allison. The third and most important aim of this thesis was to strengthen and endorse a regulative reading of Kant's radical evil, which suggests that Kant's motive for the doctrine of radical evil was not to claim that the universality of human radical evil is an actual state of affairs, but rather that we ought to presuppose human beings as evil. I have

strengthened the case for the regulative reading by highlighting the crucial role of the inscrutability thesis, providing a distinction of the cognitive attitudes of belief and presupposition, and offering a practical insight that we can gain from the regulative reading. Then, based on an analogical objection to the argument of the regulative reading, I shed light on the mindset for moral growth that the regulative reading encourages which may benefit the moral agent. Through a Kantian theory of evil, we may learn through introspection that there is a possibility of evil in each one of us, a sapling that can grow into monstrous moral perversity if we do not heed to the moral disciplinarian within. This presupposition elicits two beneficial attitudes from us as moral agents. Firstly, we can acknowledge our susceptibility to self-deception and make stern precautions to avoid it; secondly, we can be more lenient and compassionate with others, by being wholeheartedly humble through the recognition of the universal weakness of human willpower and the immense difficulty in the consistent commitment to becoming a better person.

Bibliography

Allison, Henry. “Ethics, Evil, and Anthropology: Remarks on Allen Wood.” Ethics 111 (2001): 594–613.

———. *Kant’s Theory of Freedom*. Cambridge: Cambridge University Press, 1990.

———. “On the Very Idea of a Propensity to Evil.” *The Journal of Value Inquiry* 36 (2002): 337–48.

Anderson-Gold, Sharon. “God and Community: An Inquiry into the Religious Implications of the Highest Good.” in *Kant’s Philosophy of Religion Reconsidered,* edited by Philip Rossi and Michael Wreen, 113–31. Bloomington: Indiana University Press, 1991.

———, and Pablo Muchnik, eds. *Kant’s Anatomy of Evil*. Cambridge: Cambridge University Press, 2010.

Bernstein, Richard J. “Radical Evil: Kant at war with himself.” *Rethinking Evil* (2001): 55–85.

Bratman, Michael E. “Practical reasoning and acceptance in a context.” *Mind* 101.401 (1992): 1-15.

Brooks, David. *The Road to Character*. Random House, 2015.

Chignell, Andrew. "Belief in Kant." *The Philosophical Review* 116.3 (2007): 323-360.

Feinberg, Joel. “Psychological egoism.” In Joel Feinberg & Russ Shafer-Landau (eds.), *Reason and Responsibility,* 10th ed. (1999): 493–505.

Friedman, Michael. “Regulative and constitutive.” *The Southern Journal of Philosophy* 30. S1 (1992): 73–102.

Kant, Immanuel. *Critique of Practical Reason.* Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 2015.

———. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor and Jens Timmermann. Cambridge: Cambridge University Press, 2012.

———. *Metaphysics of Morals.* Translated by Mary Gregor. Vol. 6. Cambridge: Cambridge University Press, 1999.

———. *Religion Within the Boundaries of Mere Reason and other writings*. Cambridge University Press, 2012.

Kohl, Markus. "Radical Evil as a Regulative Idea." *Journal of the History of Philosophy* 55.4 (2017): 641–673.

Louden, Robert L. "Evil Everywhere: The Ordinariness of Kantian Radical Evil." In Anderson-Gold and Muchnik, *Kant's Anatomy of Evil* (2009): 93–115.

Morgan, Seiriol. "The Missing Formal Proof of Humanity's Radical Evil in Kant's Religion." *Philosophical Review* 114 (2005): 63–114.

Muchnik, Pablo. "An Alternative Proof of the Universal Propensity to Evil." In Anderson-Gold and Muchnik, *Kant's Anatomy of Evil* (2009): 116–43.

O'Connor, Daniel. "Good and Evil Disposition." *Kant-Studien*, 76.1-4 (1985): 288–302.

Wood, Allen. "Kant and the Intelligibility of Evil." In Anderson-Gold and Muchnik, *Kant's Anatomy of Evil* (2009): 116–43.

———. *Kant's Ethical Thought*. Cambridge: Cambridge University Press, 1999.

국문초록

칸트의 근본악에 대한 규제적 해석

서울대학교 대학원

철학과 서양철학전공

백 서 원

임마누엘 칸트는 그의 저서 <이성의 한계 안에서의 종교>에서 '모든 인간은 본성적으로 악하다'는 주장을 개진한다. '근본악 테제'로 알려진 이러한 주장은 그 해석을 어렵게 하는 여러 요인들로 인해 그간 주된 연구의 대상이 아니었으나, 최근에 이르러 근본악 테제를 칸트의 전체 도덕철학적 기획 속에서 중요한 의미를 갖는 것으로 이해하려는 시도들이 이루어지고 있다. 본 논문은 이러한 시도의 일환으로, 근본악 테제에 대한 한 가지 해석법을 옹호하고자 한다. 이를 위해 제 1 장에서 예비적 작업으로 칸트의 근본악 테제를 이루는 핵심 논의들을 정리하고, 근본악 테제를 일관적으로 해석하기 위해 해결되어야 할 난제들을 확인한다. 제 2 장에서는 그러한 해석상의 문제들에 비추어 기존의 두 가지 해석들을 간략하게 비판한 뒤, 제 3 장에서 본 논문의 주된 목표로서 근본악 테제에 대한 규제적 해석(regulative reading)을 옹호하고, 이를 강화하는 데에 기여하는 작업을 시도한다.

칸트의 근본악 테제는 텍스트 상에 나타나는 다양한 긴장들로 인해 이를 정합적으로 이해하는 데 많은 어려움이 따르는데, 기존의 논의들에서는 일관적인 해석에 필수적인 핵심적 쟁점들이 무엇인지에

대한 합의의 부재로 인해 문제들의 일부분만을 해소하는 데에 그치거나 문헌해석상의 지엽적인 논쟁으로 번지는 경우가 많았다. 이를 방지하고 이론적 평가의 척도를 세우고자 본고에서는 칸트의 근본악 테제를 정합적으로 해석하기 위한 기준으로서 이전 문헌들에서 제기된 문제들 중 필수적으로 해결되어야 할 세 가지 난제를 제시하였다. 칸트의 근본악 테제에 대한 만족스러운 해석은 이상의 세 가지 난제에 충분한 응답을 제시하면서 일관적으로 근본악 이론을 설명해낼 수 있는 해석일 것이다. 제 2 장에서는 근본악 테제에 대한 기존의 대표적인 두 가지 해석법인 Allen Wood 의 인간학적 해석과 Henry Allison 의 연역적 해석이 이러한 세 가지 문제들을 해결할 수 있는지를 검토한다.

제 3 장에서는 본고에서 옹호하고자 하는 규제적 해석의 내용과 그 근거들이 구체적으로 제시된다. 규제적 해석은 근본악 테제를 제시한 칸트가 실제로 그것이 현실 세계에서 객관적 참인 사태로서 성립된다고 주장한 것인지에 대해 의심을 제기한다. 그보다는 칸트의 기존 논의에서 인간의 도덕적 삶을 위해 필수적으로 요청되는 전제들인 자유의지, 신존재, 영혼불멸 등과 같이, 인간이 도덕적 이상을 추구하고 이를 통해 도덕적 자기완성을 이루기 위한 수행에 필수적인 전제로서 근본악 테제가 요청된다고 이해하는 것이 보다 타당하다는 견해이다.

칸트는 자유의지, 신존재, 영혼불멸, 그리고 근본악 테제에 대해 서술할 때에 모두 동일한 어휘로 상정(presupposition, *Voraussetzung*)을 사용한다. 그런데 규제적 해석이 성립하기 위해서는 실천적 추론에 있어 그 테제들이 '마치 참인 것처럼' 간주하는 이러한 '상정'의 명제태도가 정확히 어떠한 인식적 상태를 지시하는지를 분명히 밝혀야만 대상을 참으로 여기는 '믿음'의 명제태도와 유의미한 구분을 할 수 있으며, 이 구분이 이루어져야만 칸트가 근본악 테제를 객관적 참으로 주장하고자 하지 않았다는 해석을 고수할 수 있다. 따라서 본고에서는

상정과 믿음의 구분을 위해 Michael Bratman 의 belief-acceptance 구분과, 칸트의 믿음 체계를 분석한 Andrew Chignell 의 연구를 응용하여 상정과 믿음의 태도가 위의 belief-acceptance/objectively sufficient assent-*Glaube* 구분과 동일하게 질적으로 구분될 수 있음을 보임으로써 규제적 해석을 강화하고자 한다.

근본악 테제의 상정이 도덕적 수행에 필수적으로 요청되는 이유는 크게 세 가지로 이해할 수 있다. 첫째, 칸트가 제시한바 인간의 악이 발현되는 세 단계들 중 보다 흔히 목격되는 두 단계들은 모두 '자기기만'을 수반한다. 둘째, 칸트는 어떤 행위자라도 그 자신의 진의를 확실하게 인식할 수 없다는 '불투명성 논제'를 주장한다. 셋째, 칸트가 제시하는 도덕적 자기완성은 인간으로서 이루기 불가능에 가까울 정도로 어려운 것이지만, 또한 실천이성을 가진 행위자로서 인간이 마땅히 추구해야 할 의무이므로 인간의 도덕적 발전이란 언제나 끊임없는 고행의 과정이다. 이상의 세 가지 준거를 종합하여 볼 때, 규제적 해석의 또다른 옹호자인 Markus Kohl 은 도덕적 자기완성의 과정에 임하고자 하는 행위자에게는 자신이 도덕적으로 악한 성향을 가지고 있다고 상정하는 것이 가장 실천적으로 이익이 된다는 논변을 제시하고, 그렇게 하지 않을 경우, 곧 행위자가 자신이 도덕적으로 선한 성향을 가지고 있다고 상정하거나 또는 자신의 현재의 도덕적 성향에 대한 판단을 유보하게 될 경우 연루되는 자기기만의 형태들을 제시한다. 즉, 근본악의 상정은 스스로의 최상위 준칙을 인식할 수 없는 행위자가 자기기만에 빠지게 될 가능성을 최소화하기 위한 안전 대책의 일종으로 이해할 수 있다.

본고에서는 Kohl 의 논지에 더하여 규제적 해석을 강화하는 데에 기여하기 위하여 크게 세 가지의 이론적 장점을 추가적으로 제시한다. 첫째, 규제적 해석은 본고의 제 1 장에서 제시한 칸트의 근본악 테제의

일관적 해석을 위한 세 가지 난제들을 모두 만족스럽게 해결할 수 있다. 둘째, 규제적 해석은 다른 해석들에서 문제가 될 법한 '공통 불투명성 논제'의 문제를 발생시키지 않는다. 칸트는 불투명성 논제를 통해 어떤 행위자도 그 자신의 최상위 준칙을 인식할 수 없다고 주장하는데, 만약 칸트가 동시에 스스로는 모든 인간의 최상위 준칙이 악하다고 주장하는 것으로 이해한다면 그 스스로 정당화되지 않은 주장을 하고 있다는 문제가 발생할 수 있다. 그러나 규제적 해석은 불투명성 논제에 충실하면서 칸트의 근본악 테제가 갖는 의미를 설명할 수 있다. 셋째, 규제적 해석은 칸트의 도덕철학의 다른 전제들을 모두 받아들이지 않더라도, 현대적인 관점에서 오만함과 자기기만을 경계하고 덕스러운 사람이 되고자 하는 행위자에게 귀감이 될 수 있다.

마지막으로 본고에서는 근본악에 대한 규제적 해석이 갖는 실천적 함의에 제기될 법한 비판을 제시하고, 이에 대한 해명을 모색한다. 비판의 논지는 과연 개인의 도덕적 수행을 위해서 근본악과 같이 극단적인 가정을 받아들일 필요가 있는가이다. 즉, 만약 도덕적 이상이 그토록 도달하기 어렵고 도덕적 성장이 힘든 길이라면, 적어도 개인이 점점 나은 사람이 되어가고 있다는 희망적인 동기부여가 필요하지 않겠냐는 문제제기이다. 도덕적 성장을 위해 근본악을 상정하는 태도가 효율적일지, 혹은 개인의 도덕적 성장을 긍정하는 낙관적 태도가 효율적일지는 결국 잘못된 자기인식으로 인한 오만과 자기기만이 주는 절망과, 완벽한 이상에 도달할 수 없는 개인의 근본적 한계로 인해 느끼는 절망 중 어느 것이 더 도덕적 행위자를 낙담시키는가의 선택의 문제로 볼 수 있다. 이에 대하여서는 도덕적 성장이라는 실천적 목표를 달성하는 데에 효과적인 방식에 대한 직관이 개인에 따라 다를 수 있음을 인정하며, 적어도 오만함과 자기기만에 빠지지 않기 위한 안전장치로써 행위자 자신의 근본적인 도덕적 한계를 상정하는 태도가 비합리적이거나 실현 불가능한 종류의 태도는 아니라는 것을 보이고자

하였다. 이상의 작업들을 통해 본고는 칸트의 근본악 테제 해석에 쟁점이 되는 문제들을 파악하고, 이를 기반으로 근본악 테제에 대한 규제적 해석을 옹호하고 이를 강화하는 데에 기여하고자 시도하였다.

주요어: 칸트(Immanuel Kant), 근본악(radical evil), 도덕적 성향(moral disposition), 최상위 준칙(supreme maxim), 불투명성 논제(inscrutability thesis), 규제적(regulative), 상정(presupposition), 자기기만(self-deception), 도덕적 성장(moral growth), 도덕적 자기완성(moral self-perfection)

학번: 2016-20081