UC Riverside
UC Riverside Electronic Theses and Dissertations

Title
Kant and the Significance of the Self

Permalink
https://escholarship.org/uc/item/9wn2b55j

Author
Morris, Courtney

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Kant and the Significance of the Self

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Courtney Allison Morris

June 2015

Dissertation Committee:
Dr. Pierre Keller, Chairperson
Dr. Jill Buroker
Dr. Maudemarie Clark
Dr. Andrews Reath
Dr. Larry Wright

Copyright by
Courtney Allison Morris
2015

The Dissertation of Courtney Allison Morris is approved:

_____

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgments

UCR is a special place to do philosophy. The professors are unusually committed to the graduate students and engage in passionate and serious dialogue not just in classrooms, but in halls, parking lots, and their own homes. I am grateful for having been a member of such a department. This dissertation took root in two pro-seminars during my first year in the program: Robin Jeshion's on Self-knowledge and Self-reference and Erich Reck's on Explanation. Those seminars were exactly the right way to begin a graduate career, if only to work with Robin and Erich who are both extraordinary teachers and scholars.

I was lucky to have a dissertation committee full of my philosophical heroes. Pierre Keller spent countless hours with me discussing Kant and everything related to Kant, which, as Pierre has shown me, is *everything*. Pierre constantly pushed me to my intellectual limits; he not only encouraged me to strengthen my existing philosophical talents, he challenged me to develop new ones. Those familiar with Larry Wright's work will recognize his influence in every chapter of this dissertation. Andy Reath is not only a formidable Kant scholar, but a careful reader of students' work. Maudemarie Clark began teaching at Riverside my second year, which was a great stroke of luck for me. Watching her do philosophy is to want to emulate it. Jill Buroker is an exacting thinker and a careful reader of Kant. Her sharp and insightful eye has been a model for me, and her comments are every writer's dream. I thank each of them for their guidance and feedback.

At my time at UCR, I have been surrounded by a supportive community of mentors and peers. Robin Jeshion was a fantastic mentor and role model. Coleen Macnamara continues to be an amazing mentor. She spent an enormous amount of time with me, poring over my documents at every stage. I owe her a lot. Samantha Matherne has been a good friend and interlocutor of all things Kant. Others I would like to thank include friends and members of the RON writing group: Justin Coates, Joe Cressotti, Megs Gendreau, Michael Goerger, Heinrik Hellwig, Bach Ho, Meredith McFadden, Ben Mitchell-Yellin, Luis Montes, John Ramsey, Patrick Ryan, Megan Stotts, and Justin White. Clinton Tolley was kind enough to add his valuable voice to my prospectus committee. I also thank the members and organizers of the UCR dissertation writer's retreat, particularly Maggie Grover, Rory Moore, and Janet O'Shea. I would also like to thank my wonderful colleagues and students at West Point, particularly the very talented students in my *Kant and 19th Century Philosophy* course.

My regular conversations about Kant with Daniel Ehrlich have helped me tremendously. I thank him for his friendship and for his enthusiastic willingness to discuss even the most obscure details of Kant's theory. Whitney Morris and Charly provided a good support system especially at the beginning of my graduate work. I also thank Cyndi Tromba. In some ways, this dissertation is the continuation of a philosophical argument we have been having for decades. I have no doubt that it settles nothing. I owe a particularly special thanks to Monique "CT" Wonderly. She has encouraged me in those inevitable dark hours of dissertation writing, has celebrated with me when things were good, and has given me invaluable feedback. Most importantly,

she has laughed with me. I cherish her friendship and marvel at her philosophical mind. Here's to *GWU*, Monique.

I owe my husband, Adam Stockman, more thanks than I can articulate without writing another dissertation. One has to be brave to begin a relationship with a dissertation writer; one has to be certifiably crazy to marry one. This document has been our constant marriage companion, and Adam has accepted, embraced, and proofread it. He has supported me, cheered me on, and made significant life changes for me. He is a real partner. I thank him for all the big stuff, but more importantly, for all the little stuff, including all the philosophical conversations along the way (especially those not about Kant).

Finally, I would like to thank my parents. One memory stands out as I put the finishing touches on this dissertation. It was late summer about nine years ago. My parents and I were sitting on their patio and I was lamenting that even though I loved philosophy and would love to pursue it as a career, going to graduate school was surely idiotic: it is difficult, there are no jobs, no money, and no security. They listened, smiled, and nodded. Then they told me to do it anyway. So I did.

For Mom and Dad

ABSTRACT OF THE DISSERTATION

Kant and the Significance of the Self

by

Courtney Allison Morris

Doctor of Philosophy, Graduate Program in Philosophy
University of California, Riverside, June 2015
Dr. Pierre Keller, Chairperson

Kant's philosophy, it is often thought, leads to an untenable picture of the self. He argues that we cannot have knowledge of the type of entity we are, but seems to contradict that conclusion by claiming we are free and morally responsible. Commentators accuse him of blindly prioritizing his ethical views, not recognizing that his metaphysical views prohibit the very picture of the self his moral theory necessitates. Sympathetic interpreters of Kant's view of the self focus on either his theoretical or ethical view, despite Kant's insistence that the two are interdependent.

I argue that Kant offers us a deeply consistent, unified, and plausible view of the self. For Kant, an individual necessarily thinks and acts as though he or she—and any other human being—is a substantial, complete entity with an identity that remains the

same over time. As this work shows, the significance of the self, for Kant, is *practical*, not theoretical.

A *theory* aims to gain knowledge of the self and examine what type of object it is and how it is constituted. Perhaps we are spiritual, immaterial souls, or alternatively, entities ultimately reducible to the brain or body. I establish that Kant thinks that such views are doomed to be either tautological or contradictory. This is because they answer the wrong question.

The right question is not how the self is constituted, but what we are committed to. Kant thinks the answer is clear: we must hold ourselves and others responsible, and we must believe that we are free and that our actions and choices matter. My interpretation reveals that Kant's true aim is to demonstrate how we can make good on these commitments without fear of being contradicted by theory. *Theorizing* about the self, according to Kant, is idle. But recognizing so makes room for one to act in accordance with one's commitments, both moral and explanatory. Furthermore, I prove that Kant thinks we must act under the idea of the self not just to hold ourselves and others responsible but to view the world as the causally unified place we do.

Table of Contents

Abbreviations

In citing the primary works of Kant, I have abbreviated titles. Citations to the *Critique of Pure Reason*, which I will also refer to as the "first *Critique*," will be given in-text with the appropriate academy page numbers (the "A/B" page numbers). Other primary works frequently cited have been identified in-text by the following abbreviations (see the Bibliography for the complete references):

Corr	*Correspondence*
C2	*Critique of Practical Reason*
C3	*Critique of Judgment*
G	*Grounding for the Metaphysics of Morals*
JL	*Jäsche Logic*, in *Lectures on Logic*
Progress	"What Real Progress has Metaphysics Made in Germany?"
Proleg	*Prolegomena to any Future Metaphysics*
R	*Notes and Fragments*

INTRODUCTION

Human being, you are such a difficult problem in your own eyes
No I am not able to grasp you.

—Alexander Pope, *Les contradictions de l'homme*[1]

Thinking about the self leads to some deeply challenging questions and puzzles. "I" represents me, but what does it refer to exactly? Does "I" refer to the same thing when I say, "I am hungry" as when I say, "I was thinking"? How is it that I can think about myself thinking? We know about ourselves in a way that others do not: a friend knows I took a summer trip because I shared pictures, but I know I took the trip without any investigation. Yet, we can also deceive ourselves. I may convince myself that I returned a lost wallet because I was honest, only to conclude later that I really did it to impress my companion. A friend praises me for something I did years ago on the assumption that I am still the same person now, yet we readily understand the claim, "I did it, but it wasn't really *me*," or the claim that one is no longer the same person after a traumatic brain-injury.

Philosophers who theorize about the self face formidable philosophical challenges. As Gareth Evans says, it is a topic that has exercised the most considerable of philosophers (1982, 205). The self is both a subject and an object: it is the thing that thinks, inquires, reasons, and acts. But it is also a thing that can be thought, inquired, and reasoned *about*. We seem to know it intimately and from the inside, yet we cannot say what it is. The very thing that philosophizes remains philosophically elusive. We might

[1] As quoted by Kant in *Anthropology*, 7:141 fn.

think it impossible to offer a consistent, comprehensive account that captures all this seemingly paradoxical phenomena.

This dissertation shows that Immanuel Kant offers such an account. Kant's theory, properly understood, is that human beings necessarily think and act under what I will call the "idea of the self." This is the idea that each human being is a complete, whole, singular, and substantial entity, with an identity that remains the same across time but that cannot be exhausted by reference to one's body or brain. Our thoughts and actions—not just about ourselves and others, but about the world—manifest a *deep* commitment to this idea, so deep that it is not until we philosophize that we articulate it. We then might think it reveals an unconditional truth about reality—i.e., we think that through philosophical analysis and reflection upon our commitments, we can discover truths about the self that are independent from those commitments. We might conclude that a human being is essentially *constituted* the very way in which we are committed to thinking of it. Kant, as this dissertation shows, thinks this is the root of our philosophical troubles concerning "the self."

Kant addresses these troubles, as I will show, by reminding us that the idea of the self—the very idea that guides us in our everyday commitments—is just that: an *idea*. It guides us necessarily, driven not only by the ways in which we see ourselves and others, but the very ways we make sense of and unify our experience. But this does not entail that it represents a metaphysical truth that has significance apart from the commitments that give the idea its grip on us; indeed, to think so is to make the mistake that leads to the very metaphysical problems that then seem intractable.

Kant's theory not only resolves the tensions that arise when thinking about the self, e.g., between our physical nature and our sense of our own freedom, it shows that without being guided by such an idea of the self, we would not even be able to see the physical world as the unified, and hence causal, place we do. Indeed, recognizing that the idea of the self is only an idea—a "regulative" idea as Kant calls it—is what allows us to unify all of our seemingly different human commitments.

That is the theory in a nutshell—the one that I will argue is the only proper reading of Kant's theory of the self. In this introduction, I address two initial challenges in understanding Kant's theory of the self. The first involves the difficulty of unifying Kant's disparate comments on it. The second is a difficulty of metaphysics: my interpretation seems to indicate that we are justified in thinking of ourselves in a certain way without actually knowing that we exist in such a way. Of course, only the dissertation as a whole fully addresses both issues; for now, I will introduce the problems and sketch their resolutions.

Section 1 of this introduction sketches the interpretive difficulties surrounding Kant's theory of the self. Section 2 presents the four major ways in which Kant speaks of the self, highlighting the difficulties of unifying them. Section 3 broadly outlines my proposal of how to do so. Section 4 addresses the most obvious objection to my proposal: that of the realist who might think it strange that our commitments can be perfectly justified without being supported by any absolute metaphysical truth. Section 5 presents an outline of the dissertation as a whole.

1. Problems With Interpreting Kant on the Self

For Kant scholars, and for philosophers concerned with the self, my claim that Kant provides a unified, intuitive, and helpful theory of the self will be a surprising claim. On the surface, Kant was not a philosopher centrally concerned with the self. He was concerned with the justification of knowledge—what he called theoretical philosophy—and the requirements for moral action—what he called practical philosophy. The self (or *subject*, *mind*, or *soul* as he sometimes calls it) plays many crucial roles in both, the problem is that several of these roles seem contradictory.

He insists, theoretically, that we cannot know what the self is. Some philosophers claim to know what the self is simply by thinking alone. They conclude that it must be a single, non-material thing that exists independently from everything else and remains the same over time. Kant offers a brilliant critique of such views and a precise diagnosis of where they go wrong. We can only have knowledge, he argues, of what is possibly given to us in space and time. But the self is not given this way. The self is always there, in the background, but always out of reach. You need it to think about it; hence it always transcends attempts at capturing it. To claim knowledge of something that transcends experience is to step outside the bounds of knowledge and make a judgment that we have no way of verifying.

Such agnosticism about the self, however, seems to belie Kant's own theoretical and practical commitments. Kant argues that to avoid stepping outside the bounds of knowledge, we need to investigate our capacity to reason, to know and be aware of its limits. But it is in our*selves*, he claims, that we encounter this capacity; thus, to gain

knowledge of our capacity to reason is to gain self-knowledge. Our awareness of ourselves as thinkers, furthermore, makes us conscious of ourselves as more than just sensible objects—it makes us aware of ourselves as *intelligible* entities. This self-awareness underlies our ability to understand the world as a unified whole and to make objective judgments. It verifies that we are more than what our sensible knowledge can capture and that we are in some way outside the spatiotemporal world. And this is exactly the type of thing we need to be to act morally: as something able to stand apart from spatiotemporal conditions and act freely. Kant relies on and defends this notion of the self unabashedly in his practical philosophy. But it seems to violate his warning that we can know nothing beyond what is presented to us in space and time.

Unsurprisingly, Kant scholars wring their hands, and philosophers concerned with the self—but who know little of Kant—quickly turn away. Those who dare tackle the incongruities between Kant's theoretical and practical philosophies often end up addressing the difficult topic of Kant's view of reality, stopping short of offering any positive, plausible view of the self that could emerge. Recently some have taken good steps to defend Kant's theory of the self in terms of what it can contribute to contemporary issues in the philosophy of mind, and many have offered interpretations of Kant's conception of the self as an agent. Unfortunately, those who have attempted to defend Kant's theoretical notion of the self have largely ignored the issues brought on by his ethical theory, and vice-versa.

Another interpretative problem is that just as with our own talk of the self, Kant uses a plethora of words associated with it: self [*Selbst*], I [*Ich*] (related to his strange

locution “the I think”), soul [*Seele*], subject [*Subjekt*], person, and mind [*Gemüt*]. Related concepts include consciousness [*Bewußtsein*], self-consciousness [*Selbstbewußtsein*], and the concept of apperception [*Apperzeption*], which Kant tells us is the faculty or capacity of the soul to be self-conscious, and which, to complicate things, is classified as either empirical or transcendental.

Kant speaks of the self as it appears (in contrast to the “self in itself”), the “I” as intelligence and thinking subject, the self as an object (B 153-6), the thinking Self (A 383), the subject’s empirical character, the subject’s intelligible character (A 539/B 567), the subject insofar as it is noumenon (A 541/B 569), the psychological idea (A 671/B 699), “I myself” as an object of an idea (A 682/B 710), the “dear self” (the self that acts from self-interest) (*G* 4:407), one’s “ego as it may be constituted in itself” (*G* 4:451), the human being who “regards himself as an intelligence” (*G* 4:457), the proper self (one’s will) (*G* 4:458), the self as a “noumenon,” and the self as a “phenomenon” (*C2* 5:6).[2]

It is easy to see why Patricia Kitcher says that “the problem with Kant’s views about the self is that he has too many of them because the self has too many roles to play in his system” (1982, 41). Moreover, it is unclear that any of these designations are *of* something, and if so, if they are of the same thing. If we follow Wilfred Sellars, we

[2] There is much talk of Kant’s “noumenal self,” but Kant’s actual locution is usually modified: “the self *insofar as* it is noumenon” (A 541/B 569, my emphasis). He does the same in the *Critique of Practical Reason* where he discusses the paradoxical requirement to “make oneself as subject of freedom a noumenon but at the same, with regard to nature, a phenomenon in one’s own empirical consciousness,” (C2 5:6), and (again relying on the “insofar” clause), “only insofar as this being is also regarded on the other side as a noumenon …,” (C2 5:48). Hence, when Kant speaks of the so-called noumenal self, it is as an aspect of something that can also be regarded as a phenomenon. I highlight this because it will have implications for how we are to interpret the so-called “noumenal self.”

should think that according to Kant, "the I which thinks is not, as such, identical with the I which runs" (1970, 345).

At some level, this messiness should not surprise us; if it is true that Kant's view illuminates real-world phenomena then we should expect as much. Kant's seemingly scattered characterizations of the self mirror our own. The key is to recognize how these roles form a unified view for Kant. This is precisely what many commentators have claimed cannot be done with Kant's theory of the self. Admittedly, the different roles Kant attributes to the self and the different characterizations he offers do not easily unite into a coherent picture. Since—as I contend—Kant's characterizations track real-life phenomena, this is not just Kant's difficulty, it is our own. If I am correct in saying that Kant offers a unified and systematic way to approach to theorizing about the self, understanding his view will illuminate ours and help us navigate the puzzles, concerns, and interest we have in thinking and talking about what we call the "self."

2. Kant's Different Characterizations of and Roles for the Self

a. The Thinking Self (in the "Transcendental Deduction")

Central to Kant's theoretical philosophy is what I will refer to as the "thinking self." This is the subject of thoughts—the "I" that thinks. This subject or "self" plays an essential role in Kant's explanation of how it is that we can have synthetic *a priori* knowledge, one of the major aims of the *Critique of Pure Reason*. Synthetic knowledge is knowledge about the world; *a priori* knowledge is knowledge that we do not gain from experience. Synthetic *a priori* knowledge, then, is knowledge about the world that is necessarily true known to be true not because experience has confirmed it (although it

certainly does). We know it to be true, Kant's argues, because it must be true in order for us to experience things at all. Hence, so-called synthetic *a priori* principles like "every change has a cause," or "substance underlies all change" are justified because they are necessary principles for the possibility of experience.

But the possibility of such knowledge—and the possible application of pure concepts of the understanding or "categories" (such as "cause" or "substance") in such knowledge—depends, Kant claims, on the capacity of a thinker to be self-conscious. An "I" that thinks must be able to recognize that it is doing the thinking—i.e., it must have the capacity to attach "I think" to its thoughts (B 131-2). [3] Thus the application of the pure concepts ultimately depends on a thinker being conscious of an "I," the representation of which remains the same across different thoughts. The capacity to represent this "I," Kant tells us, is one's consciousness of the identity of oneself (A 108, A 116), a "pure original, unchanging consciousness" (A 107). A thinker must be capable of representing itself as different from the things it thinks about, and this representation, Kant says, is a single, simple, unchanging representation.

We might think that what Kant means here is that an individual thinker must be conscious of being the same individual across his or her thoughts. We certainly have this type of consciousness. I take myself to be the same individual now as the one who wrote the previous paragraph. This is indeed a type of apperception for Kant—he calls it "empirical apperception"—but it is not what he is referring to when he claims that a

[3] Though awkward, I will sometimes continue to use the pronoun "it" for the thinking subject. "He" or "she" connotes individuality in a way that is potentially misleading. As Kant indicates himself, we do not really know *what* is doing the thinking. He refers to the "I or He, or It (the thing), which thinks" (A 346/ B 404).

thinker must have the capacity to attach "I think" to all its thoughts. This latter capacity is "transcendental apperception" and Kant claims that empirical apperception depends on it. But if the "I" of transcendental apperception does not by itself refer to an individual thinker, then what is a thinker supposedly conscious of through transcendental apperception? For now, I leave this as an open question. Here, it suffices to note the role of the self in Kant's explication of experience: the capacity of a thinker to be self-conscious is the fundamental principle of experience.[4]

In the Transcendental Deduction—the part of the *Critique of Pure Reason* in which Kant makes the above claims—Kant also discusses self-knowledge. Kant distinguishes between a "thing-in-itself" and the appearance of something. This distinction is at the heart of his idealism and figuring out what it amounts to is no easy task, but the thrust of the "Transcendental Analytic" chapter of the *Critique* is that we can only have knowledge of appearances, not things-in-themselves. This applies to the self too; we can only have knowledge of the self as it appears, not as it is in itself. This seems to require both an active thinking subject and the self as an object of thought—leading to what Kant admits is a paradox—presumably because self "knowledge" would not then consist of a thinker knowing about *itself*—as a thinker—but only about something else—its appearance—implying that self-knowledge as such is impossible (B 152-3). This

[4] He calls it the "supreme principle of all use of the understanding" (B 136).

doctrine, though a natural and straightforward consequence of his theory of knowledge, is one that Kant commentators find troublesome.[5]

b. *The Substantial Soul (the "Paralogisms")*

Another way in which Kant characterizes the self is as what I will call "the substantial soul." The substantial soul [*Seele*] is a metaphysically-loaded "I." In the chapter of the *Critique* called the Paralogisms, Kant presents rationalist-inspired arguments for the soul, inspired by philosophers such as Wolff, Baumgarten, Descartes, and Leibniz. Kant does this in order to critique their views; Kant himself argues that we cannot have knowledge of the self construed as a substantial soul.

The rational psychologists, as Kant labels them, conclude, by inferring from nothing but the phrase "I think," that the soul has certain metaphysical properties: that it is a substance, that it is simple, that it is identical over time (that it has "personality"), and that it can be known with more certainty than external objects. Descartes is the most easily understood culprit: after banishing from his thought every possible thing he could doubt he was left with the sure knowledge that "I think, I exist,"[6] on which he founded the rest of his knowledge. He concluded that he could know that he was a thinking substance and that knowledge of external objects is dubious and liable to error.

Kant's critique, in short, is that we cannot know that the soul is any of these things or that there is a soul the way the rationalists describe it. Of course, we cannot

[5] Discussions of the dubiousness of Kant's positive claims often focus on this problem—the paradox of self-knowledge—and the related problem of whether or not the categories apply to the self. See Allison 2004, 280-285; Ewing 1924, 124 ff.; Walsh 1975, 185; Paton 1971, 1:387 ff.; and Wilkerson 1976, 113.

[6] This is the way Descartes states the cogito in the Second Meditation (1998, 80).

know that there is not one either. This conclusion, on the face of it, might seem to contradict Kant's own argument in the Deduction that there must be an "I" that is simple, unitary, and identical across thoughts. This worry is assuaged by recognizing that Kant's argument in the Deduction was about how a thinker is compelled to *represent* itself in order for experience to be possible. Kant's critique in the Paralogisms can be seen as a warning against how this very requirement can lead a philosopher astray.[7] It is easy to confuse how we are required to represent something with how that thing must exist "in itself" apart from all the conditions that compel that representation. Indeed, the metaphysical mistakes that Kant criticizes in the Dialectic chapter of the *Critique* are based on confusions of this sort.

c. The Psychological Idea ("The Appendix to the Dialectic")

Despite Kant's argument that we cannot have knowledge of a substantial soul, he goes on to argue that we need to act as if we do have such knowledge. Kant claims that reasoning about the self and the world presupposes the use of what he calls ideas. An idea for Kant is contrasted with a concept. While a concept can only be significant through its application to possible experience, an idea has no such application. It does not represent any object in the world; I have a concept of a cup but an idea of God. Ideas are essential elements in all areas of Kant's philosophy, but three ideas take center stage in his theoretical philosophy, which he calls the "transcendental ideas of reason": "the idea of the cosmos, the idea of the self, and the idea of God. Although these ideas do not

[7] I agree here with Kitcher, who also makes the point that the only way to make sense of the Paralogisms is to connect it with Kant's own theory of apperception in the Analytic (1990, 182ff.). Karl Ameriks argues the opposite: that Kant's claims regarding the "I think" in the Analytic commit him to the very rationalist claims he critiques in the Paralogisms (Ameriks, 1982).

represent objects, Kant says that they still serve us by guiding the organization and extension of our empirical knowledge (A 671/B 699). Reason itself indicates that these ideas allow us to organize and extend our knowledge as much as we possibly can.

The idea of the self operates this way. It guides the organization and extension of knowledge by providing rules we are to follow in our investigation and organization of such. We ought to proceed as if the thing we call the "self" is a simple, identical substance, independent from other things in the world when we investigate certain things, for example, how the human mind operates (A 683/B 711).[8] This assumption will allow us to organize what we already know about the human mind in a unified way and will allow us to extend that knowledge to the unknown. But acting in accordance with this rule does not mean that we can claim to know a "self" that *is* a simple, identical, independent substance. That there is such a self to be found in experience is an illusion—what Kant calls a "transcendental illusion." The illusion is powerful and Kant maintains it is necessary, just as the moon necessarily appears larger over the horizon even when we know it is not (A 297/B 354).[9] The mistake is to confuse the illusion for reality, which is precisely what the rationalists do when they conclude that a thinker has a substantial soul.

---

[8] Each transcendental idea has a corresponding rule that suggests how the idea should guide our organization and extension of knowledge. Kant calls these rules schemata (A 670-1/B 698-9 and A 833/B 861). For Kant's statement of the schema that corresponds with the idea of the self, see A 672/B 700.

[9] Michelle Grier has made great strides in making sense of Kant's account of transcendental illusion and was one of the first commentators to argue that Kant's account of transcendental illusion needs to be distinguished from his critique of metaphysics (2001). She too argues that in some sense we must take transcendental illusion to be objectively real. I owe a lot to Grier's account. I do think, however, that it can be supplemented in a way that gives a better account of Kant's overall project. I will argue that we will not be able to see exactly why the illusion is necessary without recognizing how transcendental ideas guide both our reasoning and our understanding.

d. The Moral Self ("Resolution to the Third Antinomy," "The Canon," Groundwork, *Section III, and* Critique of Practical Reason*)*

The self also plays a crucial role in Kant's ethical theory. Moral action—action guided by reason alone—requires the actor to disregard any empirical conditions or desires that would determine her to act in a certain way; she must be able to stand apart from those conditions and be able to act according to reasons even when those reasons recommend actions that go against her self-interest. Kant is committed to the claim that moral action depends on the agent being free. But the agent must be free not just from determination by her desires and interests, she must be free to act under laws that she gives to herself. She must be, in Kant's language, autonomous.

Kant sometimes refers to the moral self, or the practical self, as the self "insofar as it is noumenal" (A 351/B 569 and C2 5:48). The noumenon, on Kant's theory, is contrasted with phenomenon. Something is phenomenal if it occurs or could possibly occur within space and time and noumenal if it cannot. Our very grasp of what "phenomenal" means gives us a negative sense of what "noumenal" means: it is the thought of a thing insofar as it is not a possible object of experience for us (B 307). Noumenon in the positive sense refers to an object that is actually cognized by another kind of intellect—an intellect that does not depend on the conditions of space and time for its experience. Kant's point in the first *Critique* is that we have a negative but not a positive sense of noumenon—we cannot apply our concepts to things that are not possibly given to us in experience. Concepts gain their significance only by their possible application to the given.

This distinction is at stake when Kant insists that in practical reason we must think of ourselves as "noumenal," or as intelligences. For many, it looks as though he is referring to a positive noumenal object that is the self, precisely what he said we could not do.[10] Furthermore, Kant's insistence that a moral agent can cause things to happen in the world looks like he is not only applying concepts illegitimately to a noumenal self, but applying the concept of cause to something not possibly given in experience—again, precisely what he said was illegitimate in speculative reason.

Kant insists, however, in the "Resolution to the Third Antinomy" and Section III of the *Groundwork of the Metaphysics of Morals* that one can view oneself from two perspectives: as a phenomenon and as a noumenon (A 546-7/B 574-5 and G 5: 451-2). The self "as phenomenon" is the self as it exists in the empirical world as something like an object subject to natural laws. One's behavior can be interpreted in a way that does not appeal to free choice. This type of interpretation might explain one's behavior by tracing how each action was determined by a previous state. Kant maintains, however,

[10] This is perhaps the single most emphasized target of critique against Kant's theory of the self. Kant's contemporary critic Hermann Pistorius was one of the first to emphasize the possible inconsistency here (Sassen 2000, 176-182). Another early critic was Friedrich Nietzsche, who eloquently says, "I am reminded of old Kant, who helped himself to the 'thing in itself'—another ridiculous thing!—and was punished for this when the 'categorical imperative' crept into his heart and made him stray back to 'God', 'soul', 'freedom', 'immortality', like a fox who strays back into his cage. Yet it had been *his* strength and cleverness that had *broken open that cage*!" (Nietzsche 2001, 188). Modern commentators have followed suit, including Kitcher 1984, 113; Wilkerson 1976, 192; Strawson 1975, 173-4; and Walker 1978, 133. Strawson claims that Kant's own arguments violate his own principles and that he seeks to "draw the bounds of sense from a point outside them, a point which, if they are rightly drawn, cannot exist" (Strawson 1975, 12). Strawson clearly has in mind Kant's theoretical commitments about the subject and Kant's remarks about the "noumenal" self as part of the "point [that] cannot exist" from which Kant draws the bounds of sense: Strawson remarks that the concept of the thinking being in general "shows the model shaking itself to pieces" (ibid., 174). Patricia Kitcher claims that a "noumenal, unknown self is an impossible target for moral criticism and it is at best unclear how we can know that an unknown self creates the formal characteristics of the phenomenal world" (1984, 113). For a look at some of the passages that are the source of the confusion, see the first *Critique*, A 541/B 569, A 546-7/B 574-5, A 554-5/B582-3 and G, 4:451-2, 4:457-8, and C2, 5:5-6, and 5:56-57.

that we can also view ourselves as able to stand apart from all natural laws as something that has the capacity to act freely. Thus, the claim that one's actions can be explained deterministically, according to Kant, is not incompatible with the claim that one's actions can be explained by reference to the agent as the origin of action. If we think that the former type of explanation rules out the latter, it is because we mistakenly think that the former has explained the world as it really is instead of the way we must represent it for the sake of experience. In Kant's language, it is to think we have knowledge of the world as it is "in itself," which is a view Kant calls "transcendental realism" (A 360). It is precisely the lack of such knowledge that leaves room for thinking of oneself as the origin of one's actions.

e. A Unified Picture?

In summary, Kant's "subject" or "self" has several important roles to play in his system, not all of which seem immediately compatible. Kant tells us on one hand, there are certain requirements for experience: if a thinker is to have thoughts that constitute knowledge of the world, the thinker must have the capacity to represent itself as a single thing identical across time. This "transcendental" capacity, Kant tells us, is necessary for the empirical capacity that we are more familiar with—the capacity to recognize that I am the same individual who did such and such yesterday. It might seem that Kant is making a claim about the nature of self-identity here: perhaps that what constitutes one's self-identity is an awareness or consciousness of being the same thinker.[11] Kant tells us in the Paralogisms, however, that it is a mistake to think that we can know there is some

[11] If this is true, Kant would be offering a Lockean account of self-identity (1979, 341).

underlying substance—identical across time—that is the self, and that doing so is to take illusion as reality.

Moreover, he continues to say that even though we cannot have knowledge of a substantial self that is identical across time, we should act as if we can. It is as if Kant is telling us to simultaneously believe that the moon is larger on the horizon without thinking it actually is. This recommendation is for the purposes of theoretical reason to help us organize and extend our knowledge in a certain way, but also for the sake of practical reason: moral action, after all, requires a "self" that remains identical across time that we can identify and to whom we can attribute responsibility. Kant's theoretical work maintains, however, that we can have no knowledge of anything outside of space and time, including the self. But his moral theory sure seems to indicate that we can and must have knowledge of such a self, in order to meet our moral requirements.

If there is a consistent, unified theory of the self in Kant it is not obvious what it is. For Kant scholars, the interest is with the consistency of Kant's system. Beyond this interpretative concern, however, is the question of whether Kant's view of the self can contribute to our own natural theorizing about the self like I claim it can. In the next section I will offer an overview of my proposal. Recognizing Kant's ultimate recommendations about the self will show that his own views are consistent and that they can indeed help us with our natural and intuitive worries that arise when thinking about the self.

3. The Self and Reason

It is useful to notice exactly how puzzles arise when we philosophize about the self. Assuming one is the same thinker and agent over a period of time is one of the most natural things we do. When one organizes thoughts about oneself, makes plans, or feels regret or pride, one takes for granted that one is the same individual over time. We take this to be true of others too when we wonder how another will react, praise or blame them, or interact with them in consistent ways over time. It is difficult to imagine what our experience of the world would be like without such presuppositions.

This is to notice that what Kant calls the "psychological idea," which recommends that we act and think as if the self is identical across time, operates and guides our thoughts about the self.[12] When we press on this natural presupposition and begin to theorize about it, however, it is easy to think that "I" represents a single, individual thing, independent from anything material, which remains the same across changes. It is easy to conclude that there is some bare metaphysical—nonmaterial—self that could exist apart from everything else. I can seem to strip away all the thoughts, experiences, and contingencies that comprise my identity and still be left with an "I," just as I could have complete amnesia and still have the thought "*I* have a headache."[13]

---

[12] Treating ourselves as if we are identical over time is only one of the recommendations of the psychological idea as Kant states it (A 672/B 700). The idea also recommends that we treat ourselves as if we are simple substances, able to exist independently from external objects. I focus on identity here because it is more obviously something we presuppose than the others—but treating ourselves as simple, independent substances, I will argue, is built into our experience as well. To say that we presuppose such an idea is not to say that we do so consciously or that we are aware of the psychological idea operating in our normal everyday experience. The idea and its schematization—as I interpret Kant—is an articulation of our mostly inarticulate presuppositions in experience.

[13] This type of thought-experiment is used to explain what contemporary philosophers call "immunity to error through misidentification." The idea is that there are at least some uses of the word "I" that cannot

Likewise, a sensory-deprivation tank could give me the illusion of being completely disembodied, but I could still have the thought "*I* cannot feel my body."[14] These types of thought-experiments make it look as though the self is a metaphysical object. This conclusion leads to philosophical questions that are profoundly hard to answer. If the self is some kind of disembodied ego, then how does it interact with the body? How do we individuate it? Where does it exist?

The key is to recognize precisely what has happened in the progression from the natural presupposition to the philosophical difficulty. Kant offers an exact prognosis of what has transpired: we have confused our natural presupposition that the self is a simple, identical subject, which serves us well in practice and theory, with a description of the world as it is and how it would be apart from any conditions. But the rule offers no such description. It only offers a recommendation of how we should go on together. Confusing the two is precisely the mistake the rationalists make.

In Kantian language this is to say that we have been fooled by transcendental illusion. The idea of the self—the psychological idea—offers us guidance on how to organize our thoughts about the world and the self. Because we so naturally rely on this idea we are deluded into thinking that there *is* a self that is a simple, identical, independent substance. But this is an illusion; albeit a natural and unavoidable one (A

---

fail to refer (1994, 82). Andrew Brook argues that Kant's theory anticipates the claim that some uses of "I" are immune to error through misidentification (2001, 9). I stand with Béatrice Longuenesse that Kant's theory does not anticipate such a claim (2007, 155). Although Kant's theory requires that a thinker have the capacity for self-consciousness and hence a potential awareness of an "I" behind all its thoughts, this "I" does not actually latch onto anything that we could know is identical over time.

[14] I take this from a similar example used by G. E. M. Anscombe (2004, 152).

297/B 354). As long as we treat the idea as a guiding rule rather than a description of reality, we do not do anything illegitimate. But treating the idea as if it describes something in reality is to step outside the limits of our knowledge. The question then, is why we are so tempted to think that the self is a metaphysical object and why this illusion is such a powerful one.

We already have a partial answer to this question. Presupposing that the self is a particular type of thing is built into our everyday experience—so it is partly the psychological idea and the success we have in extending and organizing our knowledge in accordance with it that gives the illusion its power. But there is something else that contributes to its force: the very way in which experience requires us to differentiate subjective thoughts from objective ones.

One remarkable aspect about our experience is that we can abstract from our own individual experiences, thoughts, and judgments, and view things from another perspective. It is our very capacity to do this that allows one to distinguish subjective thoughts from judgments that one takes to be true of the external world. I may wear rose-colored glasses, but I can see things as rose-colored without making the judgment that everything is—objectively—rose-colored. Recognizing that my subjective experience does not necessarily reflect how I ought to judge objects is to give significance to the idea that there is an *objective* judgment to be made. Making such an objective judgment requires me to abstract from the way I see things here and now and imagine what things are like from a different perspective—from the perspective of one not wearing rose-

colored glasses. Doing such requires me to abstract from something about myself and my individual perspective.

This capacity to abstract from one's subjective perspective and view things from the perspective of another thinker is, I will argue, precisely the capacity that leads us astray when we theorize about the self. Our success in abstracting this way leads us to think that we could do so completely and absolutely. "The light dove, in free flight cutting through the air the resistance of which it feels, could get the idea that it could do better in airless space" (A5/ B8). We trade on a capacity we use in everyday experience and assume that we can push this capacity to discover metaphysical truths. Since we are able to abstract from spatiotemporal contingencies about ourselves in order to make objective judgments about the world, we assume we can go further and abstract everything away, to the point where it seems as though there is a bare metaphysical self (B 427). This means that the illusion of a substantial soul is rooted in the very nature of our experience. Our very capacity to reason then, is what generates seemingly unsolvable puzzles about the self.[15] Dissolving those puzzles involves recognizing our commitments and what we are entitled to conclude theoretically from those commitments.

So far, it is not obvious how the so-called moral self fits in. It is not clear what Kant's theoretical conclusion has to do with the puzzles that arise with treating oneself as

[15] This is not to say that we cannot make great strides in speculating about the self or mind, as we do, for example, in the areas of brain-research or psychology. Indeed, such investigations often provide useful insights about the mind and the self. Such investigations and extensions of knowledge are possible precisely because we presuppose the psychological idea. Kant's point, I will argue, is a pragmatic one: such insights, useful as they are, should not be taken as philosophical truths about what the self *is*.

an agent. Treating the self as an agent reveals another tension in the natural roles we accord to the self: experientially, it seems as though there is a self—a single "I"—in control when we act. Furthermore, because I chose to do such and such and act in accordance with that choice, I am responsible for that action; alternatively, because I did not freely choose to do something but was pushed to do it anyway, I might not accept full responsibility for that action. Tension arises when this natural commitment is pressed upon, for it does not seem to fit with the way we think about the nature of causes, namely, that every event has one.[16] From the latter perspective, it looks as though an actor is always "pushed" and is thus not genuinely free to choose what he or she does. The self, on this perspective, would be just like another mechanical object in the world, despite any feeling otherwise.

As I said above, Kant's insistence in the Resolution to the Third Antinomy is that we are legitimately entitled to view ourselves from both perspectives—as something whose actions can be explained by reference to other causes and as something that is the genuine origin of action. The two are incompatible, Kant claims, only if we fall prey to the same mistake the rationalists did when they concluded that there is a substantial soul—namely, if we take either as a description of absolute reality instead of a perspective. "Ideas" again come into play here: just as we presuppose an idea of the self for the organization and extension of knowledge, we also presuppose an idea of the world—what Kant calls the cosmological idea—which recommends that we go on in a particular way for the sake of organizing and extending our knowledge about the world.

[16] This is the problem that Kant identifies in the "Third Antinomy" (A 444-51/B 472-9).

The cosmological idea in part entails that we will never reach a point in our investigations where we cannot continue to offer reasons and explanations for the way things are.[17]

Following this rule is enormously advantageous in science—indeed necessary—and we can apply the same rule to human behavior. Namely, we can explain an agent's behavior by appealing to previous causes that have determined the agent to act. The key—once more—is to recognize that this is a rule that we are compelled to follow and presuppose for the sake of unifying our knowledge about the world, not a description of how things are. Treating it as the latter indeed rules out the possibility for free action; recognizing it as a former does not. This is what leaves room for thinking of oneself as a moral self.

4. The Realist's Challenge

The major challenge to my interpretation is from a "realist," of the type Kant calls a "transcendental realist." Kant himself is a type of realist; he believes and argues for "empirical realism" (A 360), which is the view that external physical objects have their reality outside of the mind. But we might think the interpretation I have presented of his theory of the self is leveraged on an idealist distinction: a distinction between how we "really" are and how we must represent ourselves. Our interest, the challenge goes, is not with how we are justified in *thinking* about ourselves but in how we *really are*. Sure, I

[17] The principle that we can continue to offer explanations without reference to anything supersensible is the "schema" for the cosmological idea. As Kant states it: we have to pursue the conditions of the inner as well as the outer appearances of nature through an investigation that will nowhere be completed, **as if** nature were infinite in itself and without a first or supreme member—although, without denying, outside of all appearances, the merely intelligible primary grounds for them, we may never bring these grounds into connection with explanations of nature, because we are not acquainted with them at all" (A 672/B 700).

can think of myself as a responsible agent and as a thing that is able to act without regard to any of my desires and inclinations, but this does not entail that I *am* such a thing. It would be a moot point to say that we can think of ourselves as a certain type of thing if we are not that type of thing. And indeed, sometimes Kant himself strays from his "perspective" talk and implies that a moral self *is* a certain type of thing.[18]

The first step in understanding Kant's answer to the realist's challenge is to understand his claims about what we do know. We know we are the type of entities that must view ourselves in a certain type of way. We know that we are the type of entities that must view the world as causally unified and systematic. What we do not know is how we "are" or how the world "is" apart from how we must view ourselves and the world. The complaint from the realist is that this entails that we and the world could be something drastically different from the way we conceive of them.

But this complaint reveals a stubborn philosophical deafness on the part of the realist: for it assumes the existence of a reality that supposedly has a grip on us—that *matters to us* precisely because it possibly contradicts our views and actions—but that is entirely out of reach of anything we could reason about. How do we know it is entirely out of reach? Because Kant has shown us the limits of our reasoning capacity. Thus, his claim that human beings are free loses any threat of being proven wrong; to say that human beings "are" free is just to say that the threat otherwise has no significance for us.

[18] See especially C2 5:47-8. Kant states that the "… faculty of freedom … proves not only the possibility but the *reality* in beings who cognize this law as binding upon them" (my emphasis). Of course, Kant is careful to modify such claims by noting that this "reality" is merely a reality from the perspective of practical reason and that such claim does not extend our theoretical knowledge. The problem is figuring out exactly what this means.

We should not misunderstand: Kant does not claim that humans are free with a "we may as well think so" attitude. Kant does not commit here a fallacy of ignorance by claiming that since we cannot know we are not free, we are. To accuse Kant of this fallacy is to misunderstand his project. Human freedom is not a *post hoc* addition to his otherwise systematic epistemological and metaphysical project, it is, rather, the linchpin. Without it, reality—the cold hard reality that we *do* have access to—the one any realist should want to defend—would not have its significance for us as such. Kant's theory is that the possibility of empirical reality depends upon the idea of the self. We will then see that our initial thought that Kant is not primarily a philosopher of the self was misguided; Kant is almost *exclusively* a philosopher of the self.

5. The Structure of This Dissertation

In chapter 1 I offer a broad overview of Kant's revisionary metaphysics and his attempt to reorient us in regards to our metaphysical questions. I show that Kant's aim is to shift our expectations in regards to what counts as good answers to metaphysical questions. The assumption is that the answers to such questions have significance apart from the interests we have in asking them. What we need to recognize, according to Kant, is that our human purposes motivate what counts as genuinely good answers to such questions. I suggest that this interpretation of Kant illuminates a consistent way of understanding his talk of the "noumenal self," though the full significance of my claims will only become clear by the end of the dissertation.

Chapter 1 shows that Kant's preliminary conclusion is that we cannot have knowledge of what the self is, so long as we understand "knowledge" to consist of claims

that are significant apart from our human interests. Kant might have stopped with this purely skeptical claim. As chapter 1 points out, however, our human interests do not allow us to remain skeptical. Chapter 2 addresses why.

Kant says that the soul is a "transcendental illusion," meaning that it is something that we take for reality when we should not. But as recent commentators have stressed, he says it is a *necessary* illusion. This implies that even if we wanted to remain skeptical about the constitution of the soul our capacity to reason still compels us to understand it in a way that easily misleads us. In chapter 2 I offer a novel account of why this illusion is necessary. Our very explanatory practices compel us to conceive of the soul as something we can have knowledge of independently of our interests. In particular, they compel us to understand ourselves as absolute subjects. Once we recognize that this picture of the soul arises from commitments rooted in our explanatory practices, we are in a prime position to understand the *significance* of that picture. Namely, we are able to see its illusory nature. Though we cannot rid ourselves of the illusion, we can stop ourselves from being deceived by it.

Chapter 2 shows that in order not to be fooled by the transcendental illusion of the soul, we must recognize how our interests determine our conclusions. Kant has a specific name for this process of recognizing how our interests determine our conclusions: transcendental reflection. Thus, a philosopher who is taken in by transcendental illusion—as Kant argues both the rationalists and empiricists are—is a philosopher who fails to transcendentally reflect. In chapter 3 I detail the mistakes of the rationalist through this lens. I argue that in order to genuinely understand Kant's critique of rational

psychology, we must understand rational psychology as an endeavor that misidentifies the significance of its own claims. Its claims—that we are simple subjects identical over time and separate from our bodies—are correct (on Kant's own view) if we understand them as rooted in our very intellectual capacity. It is to take these claims as significant apart from this capacity that the rationalist errs.

But it is not only the rationalist who errs in regards to the self. The empiricist does too. In chapter 4 I argue that we can find in Kant a substantial argument against a materialistic view of the self, much more so than commentators have previously acknowledged. The Antinomy chapter of the *Critique* provides the clue: while it explicitly addresses questions concerning the universe, some of it is clearly relevant to what Kant thinks of the soul. In particular, I argue that on a traditional picture, the soul, from a third-personal perspective, looks like it should be both immaterial *and* material. In other words, for the transcendental realist, there is a contradiction that lurks in our reasoning about the soul. This shows first that Kant can launch an argument not only against the rationalist view of the soul but also against an empiricist one. Second, it shows how transcendental idealism is supposed to solve the contradiction that arises when we think of the soul as a transcendentally real object. This illuminates, in a way not previously seen in the literature, how we should understand Kant's positive view of the self. Namely, that we should understand Kant's view to be that the self is *non*-material.

Kant's positive view of the self is the topic of chapter 5. At this point it will be clear that truly understanding Kant's negative claims is the only way to truly understand

his positive view. His negative claims show that we should not think of the self as *something*, immaterial or material. Rather, as I will discuss in chapter 5, the self, for Kant, is an *idea* that necessarily guides humans in their actions and their thoughts, about themselves and others. In particular, I argue that humans must act under the idea of a "person as a whole." The idea of a person as a whole is related to Kant's notion of an "intelligible character" and to his comments that we must think of ourselves as noumenal objects. I will show that Kant thinks that even determinists are implicitly committed to viewing humans under the idea of a person as a whole, which is to say that they are also committed to the idea of freedom. In fact, as I show, the idea of freedom guides us not only in how we act and think about ourselves and others, it guides even our understanding of the mechanistic world. The idea of freedom truly is, for Kant, the keystone of his entire philosophical system (C2 5:3). The claim that selves are entities that necessarily act under such ideas helps us understand the proper significance of Kant's comments regarding the self in his moral philosophy, which I will also address in chapter 5. I conclude the dissertation by discussing what implications my interpretation has for Kant scholarship and for Kant's ability to contribute to contemporary debates concerning self-consciousness, self-identity, and self-reference.

THE SELF UNDER KANT'S REVOLUTION

CHAPTER 1

In this chapter I aim to show how Kant offers us a picture of the self that does justice to what is genuinely at stake when we think and ask questions about the self. These are worries or questions that arise when thinking about the meaning and value of life: questions about our own death ("will *I* survive death?"), worries that arise when circumstances go wrong ("Am *I* responsible for what happened?" or "could *I* have done something differently?"), questions that arise when physical injury threatens one's personality ("After my brain tumor is removed, will I still be *me*?"), judgments we make about other's actions ("Given the horrendous abuse the defendant suffered, can we blame *him*?") and even theoretical concerns like those we might have when we worry that perhaps our considered judgments about the world are not objective, but instead biased somehow by who we are. The questions share in common the question of what the self is. How we construe the self is not just a philosophical puzzle—we care because we have important things at stake. Kant's contributions here have gone largely unnoticed. I will show he succeeds at offering a theory of the self that captures what we really care about when we ask the question of what the self is.

The picture is a subtle one and requires us to delve into a thorny issue in Kantian philosophy—for I claim that Kant can help us precisely where most commentators think his theory falls apart. Perhaps the most familiar criticism of Kant's philosophy is that his theoretical philosophy, which covers our investigations of how the world is, and his practical philosophy, which covers ethics and how the world should be, are inconsistent.

The usual story goes like this: in the *Critique of Pure Reason*, Kant demolishes traditional metaphysics. Rationalistic philosophers before Kant, from Plato to Leibniz, claimed to know things about God, the make-up of the cosmos, and the soul. The most successful part of Kant's project in the *Critique*, the charge goes, was to show that no such knowledge is possible, a tall achievement that Kant then ruins by attempting to salvage the ideas of God, the cosmos, and the soul for the sake of morality.

The inconsistency, according to these commentators, is particularly egregious with Kant's talk of the self. His theoretical philosophy concludes we can know nothing about what the self is or what properties it has—i.e., whether it is one thing as opposed to many things, whether it is a substance that maintains its identity over time despite its changing attributes, or whether it is something that can exist on its own, separately and independently from the body. In his practical philosophy, however, he appears to turn right around and endorse a view of the self that looks to many interpreters exactly like the one he so meticulously tore down. Our moral life, Kant seems to claim, relies on our having knowledge of ourselves as free, able to cause changes in the world, and as responsible for our actions (A 539/B 567). Kant also implies that this moral self is somehow outside of space and time. This picture of the self, which founds Kant's moral system, has come to be known as the "noumenal self." It is puzzling partly because Kant insists in his theoretical philosophy that we cannot know the self to be the type of thing that can accommodate such moral aspects.

I argue that commentators have misunderstood what Kant means by his so-called "noumenal self" and that there is no inconsistency in Kant's system. A thorough

understanding of Kant's critique of metaphysics reveals that Kant's revolution in philosophy has significantly changed the nature of our questions about the self and thus what counts as good answers to those questions. While traditional metaphysical approaches have focused on the question of what the self is and what properties we can know it to have, Kant's critique shows us that we cannot answer that question—we must reinterpret it. Kant approaches metaphysical inquiry in a new way. Rather than trying to answer the question of what the self is, Kant tries to answer the question of what we are *committed to* when thinking about ourselves and our experience in the world, given the range of interests that we have as humans. Hence, Kant's talk of the self—after his critique—cannot be read in the context of traditional metaphysics. The view of the self that emerges, while ignoring the question of what the self ultimately is, satisfies what we really care about when asking that question.

I begin in section 1 by addressing Kant's technical notion of "noumenon" to see what he might mean with his talk of the noumenal self. Such talk is largely motivated by what Kant takes our interests to be when we inquire about things like the self. I discuss those interests in section 2. In section 3 I take a deeper look at the historical context to which Kant was reacting and discuss why he finds the major trends in philosophy ultimately unsatisfying. Section 4 will show how Kant's critique of metaphysics and the revolutionary method he aims to replace it with changes the nature of our metaphysical questions. This prepares us to see, in section 5, how Kant recommends we think about the self under his revolution and what he does and does not mean in referring to the self as noumenal.

1. The Problem: Noumenon

A bit of background is necessary to show why the so-called "noumenal self" has caused controversy since the earliest of Kant's critics. In the section of the first *Critique* called the "Transcendental Analytic," Kant is in part responding to Hume's conclusion that our belief in a general causal principle does not reflect a necessary law grounded in physical objects but is rather a mere habit or tendency of the mind to see two things as always connected. Thus, when we judge two things to be causally related, the relation only consists in our own association of those things, not between the objects themselves—i.e., we make a subjective judgment rather than an *objective* one. If this is true, then we are not justified in making causal judgments (Hume 1974, Section I, Part II).

Kant rightfully acknowledges, contrary to many interpretations of Hume, that Hume did not thereby doubt that our concept of cause is applicable, useful, and indispensible for experience. Rather, Hume was attempting to show something about the *origin* of our concept of causality: that we cannot know a causal judgment to be true through reasoning. It is, rather, derived exclusively from experience (C2, 5:50-51). This conclusion deeply troubled Kant because it entails that science, which essentially relies on causal judgments, cannot claim to be objective or justified by reason. Recall that whether we have the ability to be objective and unbiased in our judgments about the world is something we have a stake in figuring out.

Furthermore, Hume's conclusion entails something about the legitimate scope of our concepts like cause: if it is true that such concepts are derived from experience, then

our use of them is solely limited to experience, which means we are not justified in making judgments about things normally thought to lie outside of experience, like the objects of metaphysics (Proleg, 4: 259). Metaphysics—insofar as it is a field in philosophy that inquires about putative supersensible objects like God and the soul—is abolished (Hume 1974, Section I). Kant himself shares some of Hume's attitude toward metaphysics and agrees that concepts like cause can only be legitimately applied to experience. He also thinks, however, that we cannot simply adopt an attitude that we are delusional if we fall into metaphysical inquiry. Kant recognized that there is a way of construing things that both grants objectivity to our causal judgments and that leaves room for ideas like God and the soul (the self). [19]

The way we can do this, Kant thinks, is by hypothesizing that there is a distinction between the way objects appear to us and how they are in themselves, i.e., how they are absolutely or independently of how we think of them (B xvi). Hume, and indeed all of western philosophy, Kant thinks, has before him assumed that we have knowledge when our thoughts and perceptions about objects match how they are in themselves. Kant's project is an exploration of what happens, both to our concepts like cause and to our metaphysical inquires, if we turn this hypothesis around and suppose that we only have

[19] A lot of Kantian commentators take it that Kant's worry about Hume's problem is one solely of epistemic justification—i.e., how it is that we are justified in making knowledge claims. On this reading, Kant fully responds to Hume by showing that our judgments can be objective. It is no doubt true that this is one of Kant's aims, but I would argue that Kant was really concerned with providing an account of theoretical knowledge that *leaves room* for claims about things that go beyond our experience—like the claim that we are free. It is not clear that Kant ever genuinely doubted that we can have knowledge of the world, rather, he was deeply moved by the possibility that knowledge of the world might conclusively prove that we are not free, that morality is an illusion, and that belief in God is delusional. Kant wants to show us that we can have objective knowledge of the world and still leave room for these things.

knowledge of appearances, not things in themselves. I say more about this hypothesis in Section 4. Here I use it to show what motivates Kant's talk of "noumenon."

"Phenomenal" objects are objects of appearance, i.e., objects that are given to us within space and time to which we can legitimately apply concepts like cause and substance. "Noumenon" is what is (in some way) distinct from phenomena (A 249/B 305). Now, in order for the very concept of a phenomenal object—an object of appearance—to make sense to us, we must think of the appearance as being an appearance *of* something. Of course, since we can only have knowledge of the appearance, we cannot even say what it is an appearance of, other than the "thing in itself"; nevertheless, something must ground the distinction between the appearance and the thing it itself. Thus, we have what Kant calls a negative sense of "noumenon": the thought of something outside of the way it appears to us (A 252/B 305). The negative sense of "noumenon" is only a boundary or limit concept (A 255/B 311), that merely represents something "**insofar as it is not an object of our sensible intuition**" (B 307, Kant's emphasis). This negative concept of noumenon is crucially important for us, Kant argues, for it is what allows us to recognize when we are making illegitimate metaphysical claims.

Because of the nature of our intellect, however, and for reasons that I will discuss in further chapters, we are tempted, Kant thinks, by an ambiguity that arises from the concept of the "thing in itself." Namely, we are prone to think that because we have a negative concept of it, we thereby have a *positive* concept of it. A positive concept of noumenon, as Kant characterizes it, would be the concept of an object the way it stands

independently of our representations of it—the way that an "intellectual intuition," e.g., God, would think of it. Such a divine intellect, Kant thinks, would not have to rely on objects being passively presented to it, rather, the way it thinks about the object is completely identical with the way the object really is, since for it, there would be no difference between the object and its representation of the object.[20] We have no such luxury. But the ambiguity in our negative concept of a thing in itself thereby makes us think that when we make claims about objects, we make claims that represent absolute reality—from a God's eye point of view we might say. This is a grave mistake according to Kant (A 307-8). "[O]ur concepts of understanding, as mere forms of thought for our sensible intuition, do not reach these [objects outside the boundaries of experience] in the least; thus that which we call noumenon must be understood to be such only in a **negative** sense" (B 309, Kant's emphasis).

The problem is that for many commentators it seems that Kant himself goes on to treat at least one thing as if it were a noumenal object in the positive sense: the self. The textual evidence for this accusation is passages that center around Kant's ethical commitments, particularly his insistence that we are free. In his discussion of whether or not we are free or determined in the first *Critique*,[21] for example, Kant says that we can think of ourselves as having an "intelligible" character ("intelligible" meaning "that in an object of sense which is not itself appearance"), and that our character, "insofar as it is

[20] Kant here is not making claim about the existence of God or the way he in fact thinks if he exists. Rather, he is just trying to show us that we can at least imagine an intellect that is different from ours—one that does not depend on sensibility.

[21] See "Resolution of the cosmological idea of the totality of the derivation of occurrences in the world from their causes," normally referred to as the "Resolution to the Third Antinomy," at A 532/B 560 ff.

noumenon" does not stand under sensible conditions and can cause changes in the world (see A 538-42/B 566-70). In explaining this intelligible character, Kant says that the "human being, who is otherwise acquainted with the whole of nature solely through sense, **knows** [*erkennt*] himself also through pure apperception … he is obviously one part phenomenon, but in another part, namely in regard to certain faculties, he is a merely **intelligible object**, because the actions of this object cannot at all be ascribed to the receptivity of sensibility" (A 546-47/B 574-75, my emphasis). Kant makes a similar point in *Groundwork for the Metaphysics of Morals*: "But yet [man] must necessarily assume that beyond his own subject's constitution as composed of nothing but appearances there must be something else as basis, namely, his **ego as constituted in itself**," (G, 4:451, my emphasis). Kant seems to be saying that ethical action essentially depends upon the self *being* an object outside the spatiotemporal world ("does not stand under sensible conditions"), something that exists in the "noumenal" realm but that can nevertheless cause things in the world (recall Kant's insistence that "cause" is only a concept that can be applied to appearances) and something we can *know* to be free.

Despite Kant's continual insistence that this talk of the self does not constitute theoretical knowledge (see especially C2, 5:56-57), commentators have had a hard time believing him. The criticism takes various forms. Kant's contemporary, Pistorius, for example, argues that Kant's explanation of how we can claim objectivity for our judgments cannot be maintained if we take seriously his claim that we can know nothing of the self: "if our own person does not have certainty, then the idea of persistence in external objects will be all the more so [uncertain]" (Sassen 2000, 180). Others say that

Kant should have stood firmly by his claim that we cannot know the self as a noumenal object and fully embraced the conclusion that follows, namely, that such a self as a noumenal object and the ethical system it grounds is unfeasible. T. E. Wilkerson, for example, claims that because "Kant's ethical views rest squarely on noumenalism positively interpreted" (1976, 192), Kant ends up with an overly rigid system of ethics motivated by an "illicit delight of a supersensible world" (ibid., 193). The implication is that if Kant followed through with his conclusion that any positive judgments about a "noumenal" realm are senseless, he would have adopted an ethics grounded in the natural world instead of "pure" reason, as Kant claims to do.[22] Patricia Kitcher doubts it is even possible to ground any ethical system on a "noumenal" self: a "noumenal, unknown self is an impossible target for moral criticism and it is at best unclear how we can know that an unknown self creates the formal characteristics of the phenomenal world" (1984, 113). In his highly influential commentary *The Bounds of Sense*, Peter Strawson claims that Kant's own arguments violate his own principles and that he seeks to "draw the bounds of sense from a point outside of them, a point which, if they are rightly drawn, cannot exist" (1975, 12).

The criticisms share in common the claim that Kant's system as a whole--which includes both his theoretical philosophy and his practical philosophy—is unfeasible. If our theorizing about the world entails that we cannot make metaphysical claims about the self, which is Kant's theoretical conclusion, and if our system of morality requires that

[22] This is reminiscent of Nietzsche's criticism, who eloquently says, "I am reminded of old Kant, who helped himself to the 'thing in itself'—another ridiculous thing!—and was punished for this when the 'categorical imperative' crept into his heart and made him stray back to 'God', 'soul', 'freedom', 'immortality', like a fox who strays back into his cage. Yet it had been *his* strength and cleverness that had *broken open that cage*!"(2001, 188).

we make precisely those forbidden metaphysical claims, we are in a genuine bind. This is more than an interpretive puzzle. It is a puzzle that captures what is at stake for us as thinkers and agents. It would be deeply unsettling for us to conclude that whatever is required for moral action forces us to be inconsistent with our best empirical accounts of the world. Having to rely on metaphysical and theoretical claims about ourselves for the sake of morality that are, by our own lights, theoretically unjustifiable, would profoundly affect our understanding of ourselves and our place in the world.

My claim is that those who accuse Kant of making strong claims about the noumenal realm, including his indication that we need a robust conception of the self to ground moral concepts like freedom, are taking Kant to remain firmly within the metaphysical tradition that he critiques. Kant's critique of that tradition reorients us: it changes the questions we ask about the self (and other putative metaphysical objects, like God and the universe as a whole), thereby changing what counts as good answers to those questions.[23] Kant's talk of the self after his critique takes on a different significance than traditional metaphysics would have it; whereas the latter has naturally focused on the question of what the self *is* and what properties we can know it to have, Kant ignores this question and focuses on *how we must think* about the self given our commitments and interests as humans. The claim that we must think of the self as partly "noumenal" should

[23] I choose the word "reorient" deliberately as an allusion to Plato. While the word "reorientation" is often used to describe how Plato suggests one gains knowledge of the forms, ("the instrument with which each learns is like an eye that cannot be turned around from darkness to light without turning the whole body. This instrument cannot be turned around from that which is coming into being without turning the whole soul until it is able to study that which is and the brightest thing that is, namely, the one we call the good" (1992, 518c). Kant's critique "reorients" us not to gain true knowledge of ideas, but to help us recognize the illusory nature of some of our metaphysical claims and how we can put such ideas to legitimate and proper use.

not be taken as a claim that expresses theoretical knowledge about a metaphysical object. The way we treat and think about the self, post-critique, is not justified in that way—rather, it is justified practically, i.e., in our actions and how they entail certain commitments. Because these commitments are deeply rooted in our interests, I will now discuss exactly what Kant takes those to be.

2. Our Interest in Metaphysics

"Plato noted very well," Kant says, that our "reason naturally exalts itself to cognitions that go much too far for any object that experience can give ever to be congruent, but that nonetheless have their reality and are by no means merely figments of the brain" (A 314/B 370-71). Although we will see that Kant has complaints against Plato, he agrees with him that our ability to reason naturally pushes us to think beyond our sensible experience—and according to Kant, it is because we have a stake in doing so—ends, goals, and purposes—that compel us to.

For one, we are creatures who ask for and give explanations. The very notion of offering or wanting an explanation presupposes a great deal of knowledge about the world—it is to assume that a vast number of conditions are met in order for something particular to call out as needing an explanation.[24] Because so much has to make sense for us to even get in the game of explaining, we can regress in our explanations—we can continue to ask "why?"[25] She swallowed the bird to catch the spider, but why did she swallow the spider? To catch the fly. "I do not know why she swallowed the fly," is

[24] I owe debt to Larry Wright here.

[25] Though most of the time we do not. Context, understanding, and audience usually give us a pretty good idea of when to stop.

neither a satisfying nor sufficient answer in this game—for we are compelled to think that that too must have an explanation; otherwise, all the conditions that have to be true in order for *any* explanation to arise would themselves ultimately lack a foundation.[26] Kant thinks this goal of wanting good, comprehensive explanations is what partly what pushes us to ask questions like "what is the soul?" and we will also see in chapters two and three that it pulls us into two quite different directions: one in which explanations must stop somewhere—so we actually succeed in offering a complete and comprehensive explanation—and one in which explanations must regress infinitely—because everything must have an explanation. Kant calls our interests in explanation our "speculative interests" (A 466-67/B 494-95).

Importantly, we are also creatures who act. The practice of giving and asking for explanations applies here too, of course, but does not completely capture what is at stake for us as agents. We are deeply invested in thinking that we can, for better or worse, make things happen, and that our actions have consequences and significance: that it *means* something that I chose not to lie, that *I* was the cause of her downfall, that had I *not* pulled the child back, he would have been hit by the car. Our certainty in everyday statements like this implies a profound commitment to the idea that the world cooperates with our intentions; we often genuinely believe that had we not done such and such the world would have been different. Though subjunctive claims like this are notoriously difficult to prove philosophically ("how could we possibly *know* the world would have been different?" a philosopher may ask), our actions, thought, and talk nevertheless

[26] See A 222-23/B279-80.

reflect such commitments. Of course, we also understand that the world does not always cooperate with our actions; that the most well-intentioned actions can have disastrous results; that our plans are sometimes foiled; and that bad things happen to good people. This understanding leads us to have a genuine stake in the answers to questions like whether we will survive death or whether there is a God, because the answers to such questions will make a concrete difference in how we live and what we think we are entitled to hope for.[27] Kant calls these our "practical interests" (A 466/B 494).

Ultimately, our practical interests and speculative interests are captured by the question of *who we are* ("What is man?" as Kant puts it), which subsumes three more particular questions: "What can I know?" "What should I do?" "What may I hope?"[28] The first question in particular further motivates us to ask metaphysical questions, the aim of which, according to Kant, is to progress from knowledge of the sensible to knowledge of the supersensible (Progress, 20: 316). The "super-sensible" refers to putative objects normally thought to lie outside our sensible experience, including the soul (the self), the universe as a whole, and God. Does the soul exist? If so, what type of thing is it? It is immortal and separate from the body? Did the universe have a beginning in space and time? What is the universe fundamentally made of? Are our actions

[27] Indeed, this understanding explains why we are ultimately guided by what Kant calls the idea of the "highest good," which is the idea of a world in which "happiness is distributed in exact proportion to morality" (C2, 5:110-11), in other words, a world in which one's happiness is not an accident but a reflection of one's virtue, a world that reflects an affinity with one's good will. I discuss this idea more in section 5 of this chapter.

[28] See A 805/B 833 for the latter three. For the first, "What is man?" see the *Jäsche Logic*, 25. There, Kant says, "Fundamentally, however, we could reckon all of this as anthropology, because the first three questions relate to the last one ["What is man?"]. Heidegger uses this as evidence that an anthropology—a science of man—is the proper outcome of Kant's inquiry (1990, 142).

determined? Is there a God? Metaphysics as a discipline attempts to give philosophically articulate answers—answers that meet logical and critical standards not explicit in our ordinary intuitive answers to such questions. But it is crucial to keep in mind Kant's reminder that such questions and their answers are rooted in common human understanding; metaphysics is not the burden or luxury of professional philosophers, it is a *human* endeavor.[29]

Despite this, Kant thinks we have not succeeded in offering conclusive answers to our questions about the soul, the universe as a whole, or God. This is not to deny that plenty of philosophers have offered coherent arguments for the existence of God, or the simplicity of the soul, or for the view that we are free and not determined. Such answers are not conclusive, according to Kant, because experience cannot validate any such claims—since the objects of inquiry are super-sensible, our sensible experience can no longer be the touchstone of truth. We must try to answer such questions through reason alone; but reason alone offers us no way to adjudicate between two perfectly valid but contradictory arguments about such things, as we can give, for example, for free will and determinism. One can offer a valid argument that we are absolutely free and not determined. One can also offer a valid argument that we are determined and not free. We cannot rely on our capacity to reason to solve the dilemma—for it is our very capacity to reason that drives both arguments.

[29] See B xxxii-xxxiii. Kant makes the same point in his ethical work, where he claims that we must begin with common-sense intuition of morality and go from there to a metaphysics of morality. This is why Section 1 of the *Groundwork for the Metaphysics of Morals* is called "Transition from common rational to philosophic moral cognition" (G, 4:393).

This is why Kant compares the metaphysician to Sisyphus, who rolls a boulder up a hill each day only to have it roll back down the next (Progress, 20: 259-60). This pessimism reflects much of Kant says about the field of metaphysics before his so-called revolution: that despite continued failures to answer metaphysical questions we still attempt to answer them, imagining that *this* time we will be successful. It would be in vain to try and stop, since, as we have already seen, the "interest that reason takes in the subject [is] the most ardent that can be entertained" (ibid.). We cannot simply walk away; our interests tie us indelibly to the search.

Moreover, we should remember what Kant says of Plato: that even though reason "exalts itself" and pushes us into thinking of things far beyond our actual experience, we are not delusional: these cognitions "nonetheless have their reality" (see again A 314/B 370-71). Of course, the reality of which Kant speaks here has to be pinned down. The key to understanding how it is that the self "has its reality" is to see how Kant's method reorients the field of metaphysical inquiry. First, we have to understand exactly where Kant thinks traditional approaches have erred.

3. Kant's Analysis of Traditional Approaches to Metaphysical Questions

Kant divides metaphysicians into empiricists and the rationalists. Though Kant himself admits of more complex distinctions among particular philosophers, the distinction captures the major issues at stake in metaphysics: what the philosopher takes to be constitutive of reality—sensible objects or ideas—and what the philosopher takes to be the origin of concepts and ideas—experience or reason. Neither view has been able to make progress in metaphysics, Kant argues. At the beginning of his critical project Kant

surmises that this failure is rooted in the faulty assumption that the truths we seek represent objects as they are absolutely, independently of us. Kant famously upends this assumption, which I will say more about in section 4. Importantly, however, he does not dismiss either empiricism or rationalism on a whole-scale level; each pinpoints something right about the way we think of the world and ourselves in it. Here, I will highlight the specific approaches and what Kant finds useful in them.

In the very last chapter of the first *Critique*, "The history of pure reason," Kant tersely summarizes the "differences of the ideas which occasioned the chief revolutions" in the history of metaphysics (A 853/B 881). One difference concerns what counts as an object of cognition, i.e., what it is that we are really thinking about when we think about objects (the "object of all our rational cognitions," as Kant puts it). The rationalists, or the "intellectualists," as Kant calls them in this paragraph, "said that in the senses there was nothing but semblance, and that only the understanding cognizes what is true." The idealism of Plato is the most obvious candidate for this view and indeed, Kant names Plato as the philosopher behind the "thesis" arguments concerning the nature of the cosmos (A 471/B 499). Kant's remarks on Plato reveal a deep attraction to Plato's theory of ideas and, arguably, Kant's notion of a divine intellect (the "intellectual intuition" that grasps objects as they are in themselves) is rooted in Plato. To see exactly how Kant's own approach reorients us, it will be good to take a look at what he borrows and what he dismisses from Plato's approach.

I have already hinted that Kant thinks Plato was onto something by recognizing that our capacity to reason compels us to think beyond our experience of things that

cannot possibly be matched by anything in our experience. For Plato, because the sensible world constantly changes, we cannot have knowledge of it. Knowledge, as opposed to opinion, comes only by knowing the forms, which are simple, non-spatial, non-temporal, eternal, and unchanging ideas in which particular sensible objects participate for their qualities (a particular beautiful thing "participates" in the form of the Beautiful). We know of the forms through recollection, Plato tells us, from a previous life with the divine. Thus, the object of knowledge is an idea, illuminated itself through what Plato calls the form of the good. [30]

Sensible objects can only approximate or serve as copies for the forms in which they participate; it is not our senses, but *nous*—intellect or reason (sometimes translated as "understanding")—that grasps these ideas. Kant understands Plato to mean that such objects, which can only be grasped through *nous*, are archetypes (*idea archetypa*) in the divine understanding (R 6050). In other words, once we have grasped an object through *nous*, we are grasping it in the way God does; indeed, grasping an object this way would be to understand the object as a noumenon in the positive sense. Recall that Kant thinks we simply do not have this type of intellect. Our intellect is "discursive," meaning it depends on both sensible and intellectual contribution (A 68/B 93).

Nevertheless, Plato's use of ideas, Kant thinks, goes a good distance in capturing our practical interests. According to Kant, things we are committed to in our moral life,

[30] Plato's distrust of our senses is evidenced in Plato, 2002a, 65. For Plato's claim that the forms are simple and unchanging, see Plato 1992, 576a and Plato 2002, 76d, 78d, and 79d; for his distinction between knowledge and opinion, see Plato 1992, 476d-478, 506-11; Plato 2002b, 98ff.; and Plato 200-, 51d ff. For the "participation" talk, see Plato 1992, 476 c-d. For Plato's theory of knowledge as recollection, see Plato 2002b, 81ff.; and Plato 2002a, 72e ff. For his discussion of the form of the good, see Plato 1992, 505-11.

such as freedom, depend on modeling ourselves according to ideas that cannot be derived from experience, like the idea of virtue:

> Whoever would draw the concepts of virtue from experience, whoever would make what can at best serve as an example for imperfect illustration into a model for a source of cognition (as many have actually done), would make of virtue an **ambiguous non-entity**, changeable with time and circumstances, useless for any sort of rule. On the contrary, we are all aware that when someone is represented as a model of virtue, we always have the true original in our own mind alone, with which we compare this alleged model and according to which alone we estimate it. (A 315/B 372, my emphasis)

Though Kant goes on to say that he does not want to treat this idea as an "archetype" or "hypostasize" it like Plato did, he emphasizes that our moral lives—and the very possibility of ascribing moral worth to any action—essentially depend on it.[31]

Despite their practical use, however, ideas, insofar as they are taken to be constitutive of reality, present major problems that Kant wants to avoid. First, as Kant notes, because Platonic ideas are absolutely separate and independent of sensible experience, we have no way of verifying their objective reality since "objectively real" concepts "pertain to possible things" (A 221/B 268). A Platonic form or idea cannot possibly show up in our experience. All we get in experience is copies or ectypes of ideas; sensible experience is for Plato, as Kant says, "nothing but semblance" (A 854/B 882). This is closely related to another problem, which is the fact that while Platonic ideas serve our practical interests well, they stop short of serving our *speculative* interests, particularly the interest we have in explaining the world through scientific investigation. This has to do with the relationship that sensible objects have with ideas.

[31] See also A 466/B 494. Kant thinks that morality is not the only realm in which Platonic ideas have good use. Kant also encourages Plato's application of ideas to objects of nature in the form of teleology (A 317/B 374).

Plato's claim that they "participate" in the forms idealizes particular things to the point that elides their specific differences and the unique context in which they appear. Idealism of this type allows us to "indulge in ideal explanations of natural appearances, and to neglect the physical investigation of them" (A 472/B 500). True objects of knowledge for Plato are separate and independent from our sensible experience.

Aristotle, Kant thinks, appropriates Plato's notion of an idea in a way that does not make the mistake of making it constitutive of reality. Under Aristotle, ideas, or "concepts of the understanding" lose their status as ultimate, independent, reality-bearers existing in a separate realm, and are instead understood as being derived from and grounded in experience (A 313/B 370). This makes Aristotle an empiricist according to Kant, though not a true empiricist, since Kant thinks he made judgments beyond what experience could ever verify.[32] Epicurus is the model empiricist according to Kant, since he was a materialist who made no pretensions to know anything outside of what could be verified through our sensible experience (A 471/B 500). Though an empiricist asserts that "reality is in the objects of the senses alone, and that everything else is imagination," they do not deny that we use concepts that are rooted in our intellect and not our senses; these concepts, however, are only logical for them (A 853-54/B 881-82).

The empiricist gets something right. They can give answers to metaphysical questions about the universe, for example, in a way that allows us to do some justice to our empirical investigation of the world in a way that Platonic idealism cannot. We have

[32] Kant makes the same accusation towards John Locke, claiming he is an especially noteworthy case of inconsistency. Locke wanted to base all knowledge in experience but goes on to assert that we can prove the existence of God—something far beyond what could ever be proven in experience (A 854/B 882).

to assume, for the sake of science, that anything of significance can be accessed and measured. Science is built on the idea that supernatural, immeasurable things have no influence on our investigations. Knowledge comes through what we can verify in our experience, not through accessing other-worldly forms, but by trial and error, guesswork, and testing. In this respect Kant thinks that the empiricist is able to fulfill at least some of our speculative interests in empirical explanation (A 468/B 496).[33]

Of course, just as Platonic rationalism made a mess of our empirical investigations, empiricism, Kant thinks, makes a mess of our practical interests. At best, it is simply silent when it comes to our practical interests, having nothing to say about our concerns as agents. At worst, it is dogmatic: along with stressing that everything we can know is within the bounds of sensible experience, it positively blocks reference to any supersensible ideas, including the ones we need in our moral or religious life (A 471/B 499). Hume's pretension to abolish metaphysics makes this mistake: he "lost sight of the positive harm that results if reason is deprived of the most important vistas, from which alone it can stake out for the will the highest goal of all the will's endeavors" (Proleg, 4: 258). Kant acknowledges that there are empiricists who make attempts at grounding moral systems on nothing but sensible-experience. Hutcheson, for examples, aims to ground morality in a "special constitution of human nature" (C2, 4:442). Recall, however, that Kant thinks this will not do. The idea of virtue, for example, if it is to have

[33] Though Kant thinks we ultimately need more for our scientific investigations than what the true empiricist can offer. While he agrees that the explanation of the physical world should be rooted in the field of possible experiences and its laws (A 468/B 496), he thinks that rationalism's reliance on ideas is essential for science too. This view shows up in the *Critique of the Power of Judgment* where Kant implies that even scientific investigation needs to be seen through the lens of teleology, implying that he thinks the empirical view by itself, at least taken in a dogmatic sense, does not even serve our speculative interests.

real purchase on guiding our actions, cannot be something that variously changes with the seasons; the idea needs to provide with us a genuine *rule* on how to act or a model of what we think good action amounts to. Reminiscent of Plato, Kant thinks that thinking of the world as completely captured by our empirical understanding of it cannot explain the necessary foundation for such ideas, since it would make them "ambiguous non-entities" (see again A 315/B 372).

Kant's challenge is to find a way to grant ideas genuine power without hypostasizing them as Plato did. The real purpose of metaphysical inquiry, on this picture, is not to trouble ourselves with metaphysical puzzles like the ones we are tempted to get into when we ask questions like "what is the self?" Rather, it is to reconcile the best empirical view of the world and the commitments we have when we scientifically investigate and theorize, with the view of ourselves that is required to make good on our common-sense moral commitments. Shifting our perspective this way is what allows Kant to legitimately talk of the self in a way that seems inconsistent if we hold onto the perspective he is replacing. I will say more about this shift of perspective now.

4. Kant's Revolution: The Experimental Method

Kant suggests that we can approach metaphysical questions in a way analogous to scientific inquiry in which we propose a hypothesis and methodologically investigate whether the empirical evidence supports that hypothesis. Kant's suggestion is that in metaphysics, we likewise ought to "seek the elements of pure reason in that **which admits of being confirmed or refuted through an experiment**" (Bxviii fn., Kant's

emphasis). Unlike scientific inquiry, there is no empirical way to test a hypothesis of metaphysics, since, by definition the objects of metaphysics are not given to us through sensibility. We need another way of testing possible hypotheses. Kant suggests one: by seeing what happens if we distinguish the traditional hypothesis of metaphysics from another possible hypothesis (Kant's own). If we can show that the traditional hypothesis leads to a contradiction, we have proof that hypothesis is incorrect. If, on the other hand, another hypothesis is able to resolve the contradiction and moreover give credence to our natural human interests, it will be confirmed (ibid.). This is Kant's proposed method of critique.

I have already hinted at the implicit hypothesis of traditional metaphysics: that we gain knowledge of the world by accurately representing absolute truths that stand independently of us, i.e. by accurately representing "things in themselves." This is the hypothesis that "cognition conforms to the object" (B xvi) and is noticeable in both rationalism and empiricism, though they construe it differently and most likely would not recognize it as a hypothesis *per se*. Both metaphysical approaches assume there is an object that constitutes fundamental reality and that we have knowledge of the supersensible, if we do at all, when we are able to accurately represent whatever that object really is. For the rationalists, fundamentally real objects are ideas; for the empiricist, they are sensible objects.

This assumption leads the rationalist and the empiricist to frustrating results. If we treat ideas as fundamental reality, our interests in scientific investigation are thwarted; if we treat sensible objects as fundamental reality, we are forced to disavow moral ideas

that as humans we cannot disavow without tremendous upheaval in the way we think about ourselves and our place in the world. Furthermore, according to Kant, if sensible objects are "thing in themselves," then Hume would be correct to say that causal judgments are merely subjective (C2, 5:53). Thus, the empiricist's reliance on this assumption puts us in the further bind of not being able to claim objectivity for our empirical judgments. The hypothesis that cognition conforms to the object, in other words, has disastrous results for our need to reconcile our interests as thinkers and agents.

Kant upends the traditional hypothesis. Instead of assuming that cognition conforms to the object, Kant proceeds as if the opposite were true: that objects conform to our cognition (B xvi). Kant frames what he means in terms of "standpoints" (B xix fn) along with the help of an analogy in which he compares his new hypothesis to Copernicus's revolutionary hypothesis that the earth circles the sun. A standpoint is simply a perspective of objects. Under the Ptolemaic view that the earth is the center of the universe, we had *one* perspective: we assumed that the way in which the orbits of stars and planets appeared to us represented their true motion. Just as that view assumes appearances are things in themselves, philosophers before Kant collapsed how an object appears with how an object is absolutely, which is to say that the rationalists and empiricists operated under a "single standpoint." Kant suggests instead a "twofold standpoint," which makes a distinction between the way something appears to us and the way it is by itself—analogous to Copernicus suggesting that if we change our perspective and think of the orbital appearances of stars and planets as just that—appearances—and admit that the way in which they appear to us might be different from how they are in

themselves and their true motion—we can solve problems like retrograde motion. Likewise, if we distinguish between how objects are in themselves and the way they appear to us, we might be able to solve problems in metaphysics (B xvi-xxii).[34]

Adopting a two-fold standpoint is what Kant is suggesting when he says we should think of objects as conforming to our cognition. It also throws light on his distinction between the thing in itself and appearance. Under the two-fold standpoint, we can have two perspectives on objects: "**on the one side** as objects of the sense and the understanding for experience, and **on the other side** as objects that are merely though at most for isolated reason striving beyond the bounds of experience" (B xvii-xix fn., Kant's emphasis). The first perspective is the perspective we have from experience in which we can legitimately apply concepts like cause, i.e., the phenomenal realm. The second is the perspective we might have when thinking about objects as they are in themselves. It is important to tread lightly here: Kant is *not* suggesting that the second constitutes a positive view of noumenal objects. His point is the merely negative one that once we make the distinction between the two perspectives, we can no longer claim that the way

[34] My interpretation here of Kant's "Copernican Revolution" is slightly different from the usual one and aligns with Norman Kemp Smith's (1918, 22-25). A lot commentators emphasize Kant's language of "objects conforming to our cognition," and take it to show that the revolutionary move was to shift from thinking of objects as things in themselves to construing them as they are meaningful *for us* (with the emphasis on the latter). While it is certainly true that Kant does this, it is not completely consistent with the analogy: Copernicus's revolution stresses an anti-anthropomorphic perspective, while on this reading, Kant stresses the opposite. More importantly, I think it emphasizes the wrong thing. The proper emphasis, I think, is the suggestion that there are *two* standpoints (which also has the advantage of making it consistent with the analogy). The disadvantage of the first interpretation is that it more strongly tempts us, I think, to collapse the two-fold standpoint into one and treat what Kant says post-critique as a single-standpoint perspective. That makes it all too easy to think that Kant is making claims about metaphysical reality when he isn't. Indeed, this seems to be what has happened as commentators have attempted to interpret his comments on the noumenal self. Emphasizing the two-fold standpoint has the advantage of reminding us that there is a *boundary* and limit to our knowledge, the upshot of which is that an empirical view of the world *leaves room for* us to rely on moral ideas without being inconsistent.

we experience objects is identical to how they really are, which is precisely at the root of the problems of rationalism and empiricism.[35]

The tremendously beneficial upshot of Kant's hypothesis is that neither the rationalist nor the empiricist can stake dogmatic claim on the truth. The rationalist cannot claim to know there is a God, that the self has certain properties, or that we are free, but importantly, the empiricist cannot claim to know otherwise. This verdict is far from skeptical or pessimistic; it leaves room for us to make good on both our speculative and practical commitments. Thus, it is crucially important to remember that this space only exists through Kant's two-fold standpoint, which is the hypothesis that there is a distinction between things in themselves and appearances or that "objects conform to our cognition."[36]

Once we adopt the two-fold standpoint, the nature of our metaphysical questions change. "What is the self?" when heard as "What is the self ultimately and really?" becomes an insignificant question. The boundary between the phenomenal and the noumenal that the two-fold standpoint draws includes the acknowledgment that empirical and rationalistic claims about the self have their place so long as we do not construe them as absolute reality. The question then becomes how it is that we ought to think of ourselves given our interests. Our interests in the self are largely practical—insofar as we

[35] Another way that Kant puts this is by calling the rationalists and empiricist's view—the single standpoint view—"transcendental realism." A transcendental realist is one who takes knowledge claims to be claims about things in themselves. Kant, on the other hand, in adopting the two-fold standpoint, is an "empirical realist," but a "transcendental idealist"(A 368-69).

[36] Kant himself admits that remembering this and taking it seriously is incredibly difficult. The very nature of our capacity to reason gives us the illusion that we have only a single standpoint. I discuss this illusion and its source in great detail in further chapters.

care about psychological questions about what constitutes our personalities and identities, we do so because we care about the practical consequences of such theories. The two-fold standpoint leaves room for our interests precisely because it clearly delineates the limits of our knowledge. "I had to deny **knowledge** in order to make room for **faith**" as Kant says (B xxx). This means that if we are to follow Kant, that space can be filled in a way that does not betray our commitment not to make positive claims about a noumenal realm, but rather in a way that allows us to genuinely follow through and make good on our commitments as agents. These commitments are not justified by appealing to what the self really is, i.e., by appealing to a theoretical argument about the absolute reality of it. Rather, they are justified because we have need to think of ourselves and others in a way that cannot be fully captured by a straight empirical story. Keeping this in mind, we can understand Kant's talk of the "noumenal" self in a way not subject to the charges of inconsistency.

5. The Significance of the Self as Noumenal

Kant thinks we are required to think of ourselves as more than just phenomenal or physical objects. We certainly are subject to physical causes and we certainly are physical beings, but we must think of ourselves as more than that. This is clear when we inquire about what is necessary for us to be thinkers; the very idea of a "thinker" requires a distinction between the subject thinking and the objects she thinks about (A 104). Kant stresses this requirement in the "Transcendental Deduction" of the first *Critique*, where he also notes that thinking requires us to assume that the "I" of the thinking subject remains identical across changing thoughts (B 133-34). I can make the judgment that "it

was raining but it has now stopped," only if I at least take it that I am the same one who thinks "it was raining," as the one who thinks "it has now stopped." This is not to say that I *know* that I am indeed the same thinker—indeed, that would be to make a positive noumenal claim—rather, it is to point out what is *logically* required to construe ourselves as thinkers who think about real—non-imaginary—objects outside of us, that is, to construe ourselves as capable of thinking *objectively*.

In thinking of ourselves this way we are already thinking of ourselves as distinguished from the phenomenal objects we think about. In this respect we might say we are already recognizing a two-fold standpoint: for thinking of object *as objects* requires us to think of ourselves as somehow outside of those objects. But notice that everything we need when we so think of ourselves can be captured in merely negative terms. We need only to recognize that who we are as thinkers cannot be completely exhausted by an empirical account of ourselves. As Kant says, "we can rightfully say that our thinking subject is not corporeal," (A 357). It might be tempting to take this further and claim that because we can make this negative claim, we are entitled to make the positive claim that we thereby *are* things that lie outside the phenomenal realm. This is what Kant accuses the rationalists of doing. Kant would remind us that we cannot know of anything outside the phenomenal. But precisely because the boundary is there, and it stops the empiricist from staking claim on absolute truth, we at least do not go wrong when we think of ourselves as something separate from phenomenon.

Of course, while this shows that Kant's conception of a thinking subject is a negative as opposed to a positive noumenal one, it does not yet address the seemingly

full-blown positive claims he makes about that same subject as an agent. Kant's theoretical arguments (e.g., that we are logically required to think of ourselves as identical over time and separate from other objects) are not meant to stand alone without the fuller context of his practical philosophy, and it is only there that we get the full import of his claim that we must think of ourselves as more than just phenomenal beings.

We can address this issue by first noticing what mileage we can get from the two-fold standpoint. One of the important consequences of adopting a two-fold standpoint is that it shows that insofar as an empiricist is dogmatic (i.e., insofar as he claims that empirical concepts and sensible objects are *all* there is and that there is positively *nothing* outside of these things), he errs. Under the two-fold standpoint, the empiricist cannot prove we are determined, or that the self is something material or something reducible to physical objects. It is important to remember what is at stake here: the reason why we care about the issue of determinism—the idea that we are completely determined by natural causes—is that on an intuitive and common-sense level in order to accept responsibility for our actions and in order to attribute responsibility to others, we have to have been genuinely free to do otherwise. Likewise, we care about whether or not the self is reducible to a physical object because if it is, then *I* die when my bodies dies. More fundamentally, because we often think that how the world *is* does not match up with how the world *should be*, we have a deep stake in knowing whether or not we can effect successful changes in the world, based on our own reasoning and intentions. Indeed, Kant argues that it is precisely these things that we have at stake that really motivates rationalists to claim to have theoretical knowledge of what the self is. The

whole reason we want a theoretical account of the self, which would be a "doctrine of the soul grounded merely on pure rational principles," is to secure "our thinking Self from the danger of materialism" (A 383).

Consider also what Kant says at the beginning of the first *Critique*, where, again appealing to our common human needs, he states that it is a "remarkable predisposition of our nature, noticeable to every human being, never to be capable of being satisfied by what is temporal (since the temporal is always insufficient for the predispositions of our whole vocation)" (B xxxii).[37] It is clear that our need to think of ourselves as more than phenomenon stems from a practical one to see ourselves and the world as conducive to our vocation. Kant does not expand here on what he thinks this vocation is, and unfortunately his explicit discussion of it is dangerously delayed to the end of the book, at which point the reader, exhausted by its difficult epistemological and metaphysical arguments, may be inclined to see it as a mere flourish instead of the book's entire *point*. I refer here to the "Canon of Pure Reason," where Kant discusses the "*ultimate end* of the pure use of our reason" (A 797/B 825, my emphasis). After reminding us of the failure of our capacity to reason to answer questions like what the soul, God, and the universe as a whole *is*, he asks, "For to what cause should the unquenchable desire to find a firm footing beyond all bounds of experience otherwise be ascribed?" (A 796/B 824) His answer is clear: the desire comes from our practical interests. "If then, these three cardinal propositions [God, immorality, and *the soul*] are not at all necessary for our

[37] The context makes clear that Kant is referring specifically to how we think of *ourselves* here. We can see this by the fact that he refers to this as a remark on the "first point," which is given previously in the text as "proof of the continuation of our soul after death drawn from the simplicity of substance."

knowing, and yet are insistently recommended to us by our reason, their importance must really concern only the **practical**" (A 799-800/B 827-28, Kant's emphasis).

As I have mentioned, one of our practical interests is to see the world as conducive to our actions. Action itself is guided by the idea that there is something the world *should* be like, including robust moral claims such as "No one should lie," and non-moral claims like "The bed should be made." Crucially, thinking of the world as conducive to our actions is precisely where our theoretical and practical interests overlap. We want an answer to the question, "If I do what I should, what may I then hope?" (A 805/B 833). This question, as Kant says, is "simultaneously practical and theoretical." This is so because we can only know what we may hope for if we know what the world allows for—i.e., how the world is. If the world is completely explained by appeal to natural causes and if we know there to be nothing outside of physical objects and the physical laws that govern them, we would be forced to acknowledge there is not much to hope for. But it the world cannot be so explained and captured, we can hope for things that deeply impact our experience: for example, that my actions will have important effects on the world, that there is a life after death, that I will be rewarded for my good actions, and that an evil person will be punished for hers.

These ideas are all guided by what Kant thinks is the ultimate idea, which guides both our theoretical investigations and our moral actions: that the world *should be* one in which each person gets the happiness he or she deserves, or as Kant puts it, a world in which "happiness is distributed in exact proportion to morality," which is an idea Kant

calls the "idea of the highest good" (C2, 5: 110-11).[38] Precisely because our theoretical account of the world does not rule out such a picture, we can think of *our* world as one in which we can act on reasons, guided by moral ideas like virtue, as opposed to being pushed to action by natural forces. Such a principle that we can act on reasons ("laws of freedom" as Kant calls them), and other such "principles of pure reason have their objective reality in their **practical use**, that is, in the moral use" (A 808/B 836, my emphasis). This picture of our world writ large—that is, the picture of how the world as it would be in complete accordance with laws of freedom—Kant calls the "moral world" (A 808/B 836).

We should not be misled, however, by Kant's use of the word "world." The moral world is an *idea*—a model for how our world could be—not a positing of a world literally outside another world construed as physical or natural. It is tempting to think that on the one hand there is the cold physical world, governed by physical laws and indifferent to human ideas, and on the other hand, another world, governed by human ideas where we somehow exist as non-physical creatures with the ability to somehow cause things to happen in it. This is precisely the wrong picture. For Kant, there is *one* world—the one we live in (hence my emphasis in the above paragraph that we can think of *our* world as one in which we can act on reasons). Kant's whole aim is to show us that our theoretical account of this world does not have to be in conflict with our practical concerns. Quite the contrary: our theoretical account stems from the exact same concerns

[38] We can see here again the influence of Plato. Just as for Plato, forms are ultimately illuminated through the idea of the good, for Kant, ideas are likewise guided by the idea of the highest good. The crucial difference is that Kant does not hypostasize any such ideas—they are not objects of ultimate reality. Rather, they merely serve to guide us in our practice.

as our practical, which is our ultimate concern to think of ourselves as living in a world in which we can make good on our moral commitments. *Ours* is a world in which we can make objective claims about physical objects *and* in which we can make genuine use of ideas that are necessary for us as agents. This is the real upshot of Kant's reply to Hume.

We can now bring this back to the puzzle of the "noumenal self." First, we should notice that despite the use of the phrase "noumenal self" in the secondary literature, Kant himself does not use it. Rather, he speaks of the self "*insofar* as it is noumenal,"[39] providing evidence for my claim that such a self does not exist in a world separate from our own. As I noted early on in this chapter, the questions we have a genuine stake in answering—"What can we know?" which captures our theoretical interests; "What should we do?" which captures our practical interests; and "What can we hope?" which captures both our theoretical and practical interests—is ultimately summarized by the question of *who we are*. Kant's talk of the self insofar as it is noumenal is designed to answer the question of who we are in a way that gives full justice and is fully harmonious with the answers we give to the other three questions.

The self, insofar as it is noumenal, is precisely the self insofar as we cannot think of it as fully captured or exhausted by an empirical account of it. We cannot think of the self as something reducible to a physical object like the brain or the body because our very capacity to think requires that we think of ourselves as separate from such objects. We do not have to think of ourselves as completely determined because our theoretical account of the world does not prove us to be so. Notice that these are *negative* claims.

[39] See A 351/B 569 and C2, 5:48.

Admittedly, Kant turns these negative claims into a positive one: that we are free and hence capable of being responsible for our actions (he even goes on to suggest that we can act as though we are immortal). The crucial point is to see that Kant is not founding these positive claims on a theoretical account of the self—it is enough to show that a theoretical view of the world does not rule such things out and that we *can* think of ourselves in a certain way. This way of thinking then gets its full justification not from any theoretical view, but from our very practices of doing so.

Recall that Pistorius criticized Kant on the ground that if we take seriously Kant's theoretical claim that we can know nothing of the self, then our claims about the objective world lose their validity. "If our own person does not have certainty, then the idea of persistence in external objects will be all the more so [uncertain]," as Pistorius said.[40] Kant tells us that the *Critique of Practical Reason* is attempt to respond to this criticism, and indeed, we can now see that he does so. For Kant the self is "real," but not in the sense that we can give a theoretical account of how it exists in itself as a determinate object. Rather, its reality is vindicated (it "nevertheless has its reality") through our commitments as both thinkers and agents. Likewise, in response to commentators who think that Kant is founding his ethics on "noumenalism positively interpreted," Kant has an answer: it is enough to show that we cannot found an ethics—and furthermore, that we cannot make good on our theoretical and practical commitments thereby satisfying our range of human interests—on a single-standpoint where we take

[40] See Section 1 above and Sassen 2000, 180.

either a dogmatic rationalist view or a dogmatic empirical view of our world as the absolute truth.

Kant's revolution in philosophy was to show that we can so secure our practical interests against the danger of materialism without having to ground our moral commitments on a theoretical account of what the self is. Indeed, no such theoretical account is available, forthcoming, or even possible. It is enough, for our moral lives, and for our lives as agents, that we are at least able to think of ourselves in a certain way and that doing so is not ruled out by our best empirical accounts of the world. Thinking of ourselves as morally responsible, free agents is then vindicated by our profound interest in doing so. This is the significance of the self, insofar as it is noumenal.

THE TRANSCENDENTAL IDEAS AS NECESSARY PRESUPPOSITIONS OF REASON

CHAPTER 2

When we try to pick out anything by itself, we find it hitched to everything else in the Universe.

—John Muir, *My First Summer in the Sierra*

In the last chapter I argued that one of Kant's major goals in the *Critique* is to change our understanding of the questions we naturally ask about ourselves, the universe, and God. Recall that we have, according to Kant, a deep craving for finality in our knowledge—we want to understand things in a way that leaves "no room … for any further **Why?**" (A 584/B 612). Our "why" questions, I claimed, can be interpreted as questions concerning the ultimate constitution of everything or as questions about what we are committed to; I argued that an important part of Kant's project is to convince us that we cannot answer the first but that we can and must answer the second.

This chapter will further illuminate the claim that Kant wants us to emphasize questions of commitment by showing that our beliefs in the self, the universe as a whole, and God, are natural and inevitable. Kant thinks that each individual will find it impossible not to believe that he or she exists as a soul, that there is a whole universe out there, and that there is a God. Such beliefs, according to Kant, are already and always woven into our experience. The essential role such ideas play for the sake of our human vocation motivates any scholarly interest in them, according to Kant. Indeed, it is not until we treat such things as having knowable natures and properties that Kant thinks such beliefs become harmful. This is how philosophers have traditionally treated them: as objects the properties and natures of which we can fully understand. The irony, Kant

thinks, is that treating these things as objects of knowledge stymies the whole *aim* of metaphysics, which is to help us see beyond our immediate experience and inform us about our vocation as human beings (B xix-xx, A 840/B 868). Kant's reorientation of metaphysics attempts to restore to it this essential task.

A crucial part of this restoration is Kant's critique of dogmatic metaphysical arguments concerning the soul, the universe, and God. I discuss such arguments and Kant's critique of them in the next two chapters. First, however, it is crucial to understand the role Kant thinks the beliefs in the self, the universe as a whole, and God, play in our everyday experience and why Kant thinks they are natural, inevitable, and necessary (A 297-98/B 353-54). That will be the topic of this chapter. Unpacking the status of the ideas of the soul, the universe, and God, is the key, I argue, to understanding why Kant thinks we inevitably take such things to exist. To use Kant's language, understanding the nature and status of these ideas is the key to understanding transcendental illusion.

Recently, good work has been done on Kant's doctrine of transcendental illusion, which is the doctrine that we inevitably think we exist as souls, that there is a whole universe out there, and that God exists. While much of the influential literature on the *Critique of Pure Reason* in the early twentieth century largely ignored what Kant says about the doctrine in favor of focusing solely on his critique of the arguments of dogmatic metaphysicians, the new work argues convincingly that we must pay attention to both. It has argued, for example, that the fallacious arguments of the dogmatic metaphysicians should be considered separate from transcendental illusion—and that while we can avoid the first, we cannot avoid the second.

This is largely due to Michelle Grier's groundbreaking work on transcendental illusion (Grier 2001).[41]

Grier argues that transcendental illusion is caused by confusing two principles, the first of which is a rule for our own understanding, "P1": "Find for the conditioned knowledge given through the understanding the unconditioned whereby its unity is brought to completion (A 308/B 364)" (2001, 199), and the second of which, "P2," is a principle that asserts objective necessity: "If the conditioned is given, the whole series of conditions, subordinated to one another—a series which is therefore itself unconditioned—is likewise given, that is, is contained in the object and its connection. (A 308/B 364)" (2001, 122). Grier interprets P1 as a rule we must follow in order for us to satisfy what Kant asserts is a demand of reason: systematic unity of thought (2001, 121). Basically, P1 says that in order to bring the highest unity to our understanding of the world, we should seek ultimate explanations. It is, however, not a law that extends to objects. It is only a law that we must follow to approach a systematic *understanding* of the world; it does not mean that *objects themselves* will actually obtain the type of unity we are attempting to find (2001, 120).

However, as Grier goes on to say, "Kant's ultimate position is that this demand for systematic unity of thought is necessarily conceived by reason as a transcendental principle which *is* objective. Indeed, Kant goes on to claim that we *cannot help but* take P1 to be objective" (2001, 121), which is to say that we move from P1 to principle P2. In

[41] Grier calls Kant's claim that transcendental illusion is inevitable and necessary the "inevitability thesis" (2001, 4ff.). She notes, correctly I believe, that the only way to make sense of this claim is to separate Kant's claims about transcendental illusion from his claims about the fallacies discussed in the Transcendental Dialectic (ibid., 9).

other words, in order to meet the demand that we unify our knowledge of the world, we are compelled to project *onto* the world a systematic unity. Transcendental illusion, then, is explained by the fact that we cannot help but project our own subjective needs onto the world as if they were objective (i.e., as if rules only meant for our own understanding apply to objects). This is why Grier often calls transcendental illusion a "propensity" to confuse the two principles (2001, 8).

Along with convincing us that we ought to keep separate the fallacies of the dogmatic metaphysicians and transcendental illusion, Grier's account also gives us a framework for understanding how transcendental illusion arises and how we are deceived by it. Grier is most certainly correct that transcendental illusion is connected with the demand for the "systematic unity of thought." Grier's account falls short, however, of explaining *why* we mistake maxims of our own subjective understanding with objectively valid truths. She surmises that "in order for P1 to have any *epistemic force*, it is *necessary* to assume it to be objectively valid" (2001, 121, first emphasis mine), without further developing what "epistemic force" may amount to.[42] Furthermore, the account fails to explain why transcendental illusion emerges only and specifically with the three particular ideas that it does. Kant tells us quite explicitly that the three transcendental ideas (the soul, the universe as a whole, and God) are not arbitrary but essentially connected to the nature of reason (A 327/B 384). This implies that there should be a

[42] According to Grier, the claim that P2 is "necessary" is "perhaps the most perplexing aspect of Kant's doctrine of transcendental illusion" (2001, 124). She goes on to say that P2 is just P1 when it is "conceived by reason in abstraction from the conditions of the understanding" (ibid.). As we will see, I agree that the transcendental ideas arise through a process of abstraction, but my interpretation accounts much better for why such ideas are necessary.

good account of exactly why our capacity to reason leads to these three ideas as opposed to any others.[43]

My account fills in these gaps by showing that we attribute objects to the ideas of the soul, the universe, and God (i.e., we confuse Grier's "P1" and "P2") because we must take such objects as given in order to understand the world as one that yields good and comprehensive explanations in answer to any given inquiry. Through this, we can see that transcendental illusion emerges with the specific ideas of the soul, the universe as a whole, and God because they are *the* necessary and basic components of an idea of a whole, integrated, and systematic experience, which is precisely what we presuppose when we ask for or offer an explanation of anything in particular.

In section one I define what it means for an idea to be a transcendental idea of reason according to Kant. Section two shows precisely what Kant thinks we are doing when we reason, particularly when we engage in explanation. This will prepare us to see the main point of section three, which is that our explanatory practices presuppose that each component of experience is part of a holistic, systematic context wherein each component shares a logical relationship with other components. Next, in section four, I will argue that the transcendental ideas of reason are necessary and fundamental parts of that holistic and systematic context. This will illuminate the legitimate use of pure reason, the topic of section five.

---

[43] Grier's account is indeed an improvement over the usual dismissal of Kant's claim that the transcendental ideas are connected essentially to the structure of reason—it goes a good distance, for instance, of illuminating why the idea of the soul might be connected with the categorical syllogism (2001, 136). Ultimately, however, she too rejects the connection as "difficult to defend" (ibid., 137). As we will see below, my interpretation provides the context that motivates Kant's seemingly difficult to defend claims on this point.

1. The Transcendental Ideas of Reason

To say that we necessarily believe in the soul, the universe as a whole, and God is to say that we must think these things exist. We cannot, however, verify their existence, and thus we cannot make objective (i.e., true or false) judgments about them. I can judge, rightly or wrongly, that a ship moves downstream: ships and streams appear to us in space and our sense of movement through time gives us a sense of direction within space (I rely on Kant's own example at A 192/B 237). I can be wrong about my judgment for a variety of reasons: perhaps I hallucinate, perhaps I misjudged the direction of the stream, or perhaps the ship really stands still. It is precisely because these objects are given to us in intuition that we can make judgments about them that others can affirm or correct. God, the universe as a whole, and the soul, on the other hand, do not show up this way. God, insofar as we think of him as immaterial and eternal, does not appear in space and time. Likewise, we cannot walk around and survey the whole universe as we do a ship. Though we are no doubt in the universe, we'd have to be *outside* the universe for it to show up as a whole object. And even if we believe that we exist as souls—spiritual objects separate from our bodies—we never run into an object that is a soul in the way we can a ship. We have only *ideas* of the soul, the universe as a whole, and God.

To say we have only an idea of something, however, is not to disparage it, or to regard it as "superfluous and nugatory" (A 328-29/B 385). Kant thinks we cannot do without ideas (his system after all, is called "transcendental idealism"). But it is important to understand precisely what Kant means by "idea." To call the soul, the universe as a whole, and God "ideas" is to first point out that we will never experience

objects congruent to them (A 328/B 385). This is one general way that Kant uses the term "idea": to refer to the representation of things not possibly given in our spatiotemporal experience. When such ideas emerge from our capacity to reason they are called "ideas of reason" (A 311/B 367)—as opposed to ideas that emerge from our capacity to *imagine* things outside of our experience, like an invisible being, which are ideas of the imagination (*CJ*, 5: 314). Moral ideas, like the idea of virtue, are ideas of reason because they create standards we rely on when reasoning about what to do but represent perfections no one will ever actually embody (A 315/B 371). "Aesthetic ideas," are the counterpart, Kant says, of ideas of reason: while ideas of reason lack spatiotemporal objects, an aesthetic idea is of an object (e.g., a painting, a symphony, a poem), but represents our imaginative thoughts about that object in a way that cannot be adequately captured through concepts (*CJ*, §49).

The ideas of the soul, the universe as a whole, and God have special status: they are ideas of reason. They are ideas because they represent objects that cannot show up as wholes within our spatiotemporal experience, they are "of reason" because, according to Kant, they emerge from our capacity to reason. But among all the ideas of reason they are unique because they alone attempt to represent something no other ideas attempt to represent: the "absolute totality of all possible experience" (*P*, 4:328, see also A 327/B 383-4). Hence Kant's official term for them, "the transcendental ideas of reason" (A 321/B 377). To understand what it means for an idea to attempt to represent the "absolute totality of all possible experience" we must first understand what we are doing, according to Kant, when we reason.

2. Our Capacity to Reason Logically

In introducing "reason" Kant first tells us that it includes our capacity to make mediated inferences—mediated because we reason about something through another principle or concept (A 299/B 355; A 330/B 386). This can go one of two ways: we can explain something in terms of something else—what Kant calls the ascending or hypothetical use of reason (and later what he calls reflective judgment) (A 331/B 388; A 646/B 674; *CJ*, 5: 179),[44] or we can infer or deduce something from something else—what Kant calls the descending or apodictic use of reason (A 331/B 356; A 646/B 674).[45]

---

[44] Kant seems to have chosen the metaphors of "ascending" and "descending" partly because of the way reasoning is usually expressed in an argument form. When we explain something we are going from the conclusion to the first premise, which is literally above it (hence "ascending"), and likewise, when we infer something we are going from the first premise to the conclusion which is below it (hence "descending"). This language, however, also taps into an overarching metaphor found across Kant's corpus in which reason—when it is an activity of unification and more obviously when it grounds morality—is considered "higher" than other capacities, e.g., *C2* 5:24. For an interesting discussion of this metaphor see Lakoff and Johnson 1980, 14-21.

Paul Guyer notices that the "regulative ideal of systematicity in empirical knowledge"—that is, the goal of organizing our empirical knowledge in a systematic (and complete) way—is assigned to the hypothetical use of reason in the first *Critique* and reassigned to the faculty of reflective judgment in the third *Critique*. His explanation for this is that in the first *Critique* Kant is focusing on the "unconditional completeness" that is a demand of *pure* reason, while in the third *Critique* Kant is focusing on the way the transcendental principles of experience apply to the sensible given, and thus, that he there has "reason to associate the ideal of systematicity with judgment rather than reason, with the task of subsumption rather than with an independent objective of completeness" (1990, 19). For my purposes in this paper it does not matter which faculty is assigned the task of organizing empirical knowledge—although I will continue to speak of reason as the faculty that does so.

[45] Jonathan Bennett objects to calling reason our capacity to make inferences, complaining that the ascending use of reason is not properly inferential at all, presumably because we are not strictly inferring anything when we hypothesize a general principle (1974, 261). However, it is clear that Kant himself used the term "inference" in a broader way than Bennett allows and that Kant surely considered both the ascending and descending use of reason "inferential." For example, when Kant discusses the ascending use of reason he says that we hypothesize a general principle and then test particular judgments to see if they fall under that principle: if they do, then the "universality of the rule is *inferred*" (A 647/B 675, my emphasis). Thus, Kant allows for *inference* to the best explanation. This is consistent with his use of the term "inference" in his logic lectures, where he says, for example, that "there are inferences, nonetheless, where we infer from the particular to the universal," admitting that "this kind of inference is completely opposed to logical rule, to be sure, but we cannot do without it, and along with it most of our cognitions would have to be abolished at the same time" (L, 287).

For reasons that will become clear, in his discussion of the transcendental ideas of reason, Kant is mainly interested in examining the first type—though we will notice some overlap between the two.[46]

Now, a reasonable explanation or deduction will not only attempt to explain or deduce something by appeal to something else, it will attempt to do so by appealing to a logical relationship between the two things. Suppose, for example, that I want to explain why Socrates is mortal. I might appeal to the fact that he is a human—and in so doing explain his mortality in terms of another concept: that of humanity (hence, the "mediated inference"). The relationship Socrates has with the concept of humanity is one of belonging: he belongs within the category of humans—which is to say, in Kant's language, that they share a "categorical" relationship (A 70/B 95; A 73/B 98). Aside from this type of relationship—Kant identifies two other possible relations: the "hypothetical" and the "disjunctive." I discuss the second more fully below—for now we can look at the hypothetical. Explaining something by appeal to its hypothetical relation with something else is just to explain it by appealing to what caused it or what it depended on, e.g., "she fell because I tripped her," or "her hay-fever kicked up because of the high pollen count."

Although we normally explain things in an informal manner—"the plant died because I forgot to water it"—if we wanted to, we could express our reasoning in a way that better articulates its form: "Lack of water causes plants to die. This plant did not get

[46] We might initially think of this distinction as the difference between reasoning (inferring, deducing, etc.) and explaining. This is a fine rough way of thinking about it, though as I said, we will see that the distinction becomes less clear-cut, particularly in regards to what Kant calls disjunctive syllogisms. For a good discussion of the relation see Larry Wright, 2002.

enough water. Therefore, it died." This is to say that we could express our reasoning through syllogisms. Kant is not saying here that we actually do or must reason through syllogisms (although it is not crazy to think so),[47] rather he is making a point that our reasoning follows particular logical patterns. If we abstract away from the content of what we reason about we see that our reasoning (which in this context includes inferring, explaining, deducing, etc.) can express a logical form: e.g., "All *A*s are *B*s, x is an A, therefore x is a B." Indeed, Kant names a syllogism for each form, giving us a categorical syllogism (the form I just schematized), a hypothetical syllogism ("If F then G, F, therefore G" or "(x)(Fx → Gx), Fx, therefore, Gx"), and a disjunctive syllogism ("Either A or B, not A, so B").

The disjunctive syllogism is slightly different from the other two. Suppose I reason as follows: "Either there is an accident on the freeway, or they are doing construction. They are not doing construction. Therefore, there is an accident." Unlike the other two forms, this seems less an example of explaining something in terms of something else and more of an example of inferring: the major premise does not explain why there is an accident in the way that the principle "all humans are mortal" explains why Socrates is mortal, or the way that the principle "lack of water causes plants to die"

[47] Kant's model seems close to the "deductive-nomological" model of explanation explicated by Carl Hempel and Paul Oppenheim, according to which a particular phenomenon is only properly explained by appeal to a set of antecedent conditions along with a statement of a general law, from which the conclusion follows via a logical deduction (1988, 12). Despite the similarities, however, there is nothing that commits Kant to thinking that we must reason through syllogisms—in fact, there is nothing that commits Kant to thinking that we must reason deductively (as opposed to inductively). I do think it is important to him that our reasoning is done *logically*—and that general principles maintain a valid or cogent connection with the particular judgments they support. As I will discuss in the next chapter, I think Kant's considered view of explanation is actually closer to what we might call a pragmatic theory of explanation—whereby our purposes and interests play a major role in what counts as explanatory.

explains why my plant is dead. Rather, the appeal to the first premise of the disjunctive syllogism explains why *I think* there is an accident—which is to say that it is an inference to the best explanation.[48] I highlight this only because Kant himself seems to classify disjunctive syllogisms under the "explanatory" ("ascending") use of reason—which initially seems strange.

We can, however, quickly transform disjunctive reasoning into something explanatory. Suppose, for example, that I want an explanation for why the traffic is backed up, and I think: "Either there is an accident or there is construction. I know they are not doing construction on this road, so there must be an accident up ahead." What we are doing when we reason this way is just concluding what is possible given what is actual. Given our understanding that some actualities cancel some possibilities out we can infer another actuality. (Of course, nothing prevents an accident *and* construction from backing up traffic, but under normal circumstances I might know one or the other to be the plausible explanation). This insight—that reasoning disjunctively expresses our general grip of possibility—will become important when I discuss the relationship this form of syllogism shares with the idea of God. For now, it is enough to note it.

We see now what Kant means when he says that part of our capacity to reason involves making mediated inferences, an important part of which is to explain things in terms of other things. That logical forms underlie this type of reasoning explains why Kant calls this use of reason, "the logical use of reason" (A 303/B 359). The logical use

[48] This is precisely what Wright says is a possible (though not necessary) difference between explanation and reasoning: that the first explains why something happened and the second explains why we have reason to think something (2002, 36.)

of reason is important, Kant thinks, because it is partly what fools philosophers into thinking that they can reason about the soul, the universe as a whole, and God in a way that yields knowledge about what type of objects they are. The next section will begin to show why.

3. Presupposing the "Totality of Conditions"

Kant notices an important difference between what we might think of as explanation (the "ascending use of reason") and inference or deduction (the "descending" use). When we explain things through reason—that is, when we explain things through another concept (in a "mediated" way)—we have to presuppose certain truths, namely, we must presuppose that there *are* conditions that would explain any given conclusion we arrive at through reason.

Examples will make this much clearer. Suppose we want an explanation for the fact that Caius is mortal (to use Kant's own example at A 322/B 378). We can search for an explanation by searching for a principle that explains his mortality—the fact that all humans are mortal. Arriving at the truth of "Caius is mortal" through reason *depends on* the truth that he is human *and* on the truth of the principle that all humans are mortal. In Kant's language, this is to say that we presuppose the truth of two "conditions" which Kant sometimes refers to as the premises of a syllogism because they can be expressed that way.

Notice, however, that the truth of these "conditions" presupposes the truth of others: there is another explanation forthcoming for why all men are mortal—perhaps that they are animals (and that all animals are mortal). Of course, the truth of these new

"conditions" presupposes the truth of yet other conditions: another explanation is forthcoming for why all animals are mortal, and so on. The conclusion that Caius is mortal presupposes the truth of the conditions all the way back, so to speak, or as Kant puts it, it presupposes the truth of the totality of conditions—the "whole series" of which must be "unconditionally true" (A 331/B 332). So when we endeavor to explain something, we must, in a sense, presuppose that an explanation is forthcoming and that this same principle continues to operate for each such forthcoming explanation. Kant refers to such continued explanations as "prosyllogisms," simply because such reasoning can be expressed as a syllogism the conclusion of which is the major premise of the one with which we began and so on (A 331/B 387-88).

Explanatory reasoning is peculiar in this way, Kant thinks: for when we deduce or infer something, we do not need to presuppose the totality of inferences that we could possibly *infer* from any given principle—for example, we do not need to presuppose that there is a complete and total number of inferences that exist and hence could be made from the principle that plants die without water. This principle could be true even if there were *no* inferences to be made—i.e., in a world that no longer has plants. We can certainly imagine all the inferences that could be made from any principle, but we do not necessarily presuppose them in order to arrive at any given truth. As Kant puts it, "the possibility of something conditioned presupposes the totality of its conditions, but not the totality of its consequences" (A 337/B 394). When we engage in explanation we have an *idea* (note the term) of a totality of conditions in the background—but when we infer we

do not necessarily have an idea of all the possible inferences we could ever make from a single principle.

Kant tells us that this peculiarity that emerges when offering or asking for explanations—the requirement that we must presuppose the truth of a totality of conditions—stems from the fact that "only under this presupposition is the judgment before us possible *a priori*" (A 331/B 388). He means that we must presuppose a totality of conditions to arrive at a conclusion through reason *alone*, as opposed to experience. Just as we can search for an explanation of the fact that Caius is mortal, we can also reason in the "other direction" so to speak, and attempt to arrive at the truth that Caius is mortal by deducing or inferring it from the principle that all men are mortal and the fact that Caius is a man. When Kant says that the judgment (e.g., "Caius is mortal") is possible *a priori* only if we assume that the totality in the series of conditions is given, he means that the only way to arrive at the truth of the claim that Caius is mortal without relying on experience (which is what makes it an *a priori* judgment), is to assume that both the major and minor premises are true. Kant admits of course, that we could conclude that Caius is mortal through experience, e.g., we could watch Caius die and subsequently judge (as opposed to deduce) that he is mortal. We could also *infer* (as opposed to deduce) that Caius is mortal through experience by recognizing that all humans we have ever observed have been mortal, and that therefore, Caius is. But if we want to arrive at the truth of this claim through reason alone, we can only do so by assuming the truth of the premises.

There is something much deeper going on, however, than our ability to make *a priori* judgments through reason. Kant is making a point about the holistic and systematic context that we must presuppose for the sake of our explanatory practices.[49] To see this, we only need to apply his point that we must assume a "totality of conditions" to an example that brings to bear the massively complicated real-life conditions in which most explanations take place. The space shuttle *Challenger* explodes mid-flight. We want to know why.

Richard Feynman explained the disaster by indicating that something called an "o-ring" did not expand properly during take-off because of the cold weather, allowing hot gases to escape from the rocket booster, causing the explosion.[50] There is an enormous number of things that might need explanation before one can understand the "o-ring" story as genuinely explanatory. One would most likely need a background in physics or engineering in order to understand *why* an o-ring not expanding would cause an explosion; indeed, most of us would begin by asking what an "o-ring" is. A good explanation is a function of purpose and audience. In some contexts, the above explanation might be sufficient, perhaps if one wants to know if the explosion was the fault of the astronauts or the fault of the NASA engineers. Nevertheless, any explanation of the disaster depends on the audience understanding a perhaps infinite number of truths regarding engineers and how they work, the atmosphere and weather systems on earth, and what space shuttles are supposed to do.

---

[49] My interpretation of Kant's theory of explanation is highly influenced by the work of Wright. See especially Wright, 1995, 565-585 and 1973, 508-509.

[50] See James Gleick, *Genius: The Life and Science of Richard Feynman* (New York: First Vintage Books, 1993), 414-428.

This all takes place in a context in which we have no trouble recognizing what is normal. The vast number of conditions that constitute normality will guide our inquiry—from the obvious fact that a shuttle blowing up in mid-flight is not normal, to the equally-obvious but considerably less-relevant fact that such events have never been explained by appeal to an alien attack, less relevant precisely because we know it is not a live option. It did not need to be announced, for instance, that the shuttle explosion was *not* caused by aliens, because nobody of sane mind suspected as much. For a contrast, consider how, after an explosion these days, we do consider terrorism as a cause to rule out. Lurking in our "o-ring" explanation is a vast wealth of important detail that is explanatorily relevant: the fact that the o-ring was made from a material that did not expand, the fact that heat rises, the fact that the shuttle was supposed to be built in a certain way, and so on. For each of these we can explain further: why certain material does not expand under certain temperatures, why heat rises, why shuttles need rocket boosters, etc. We could also direct our inquiry into why NASA management tended not to listen to the repeated warnings of the engineers about the material of the o-ring,[51] why the temperature was under 32 degrees the day of the launch, why NASA chose that day for the launch as opposed to the day before, and so on. A great deal of this was indeed relevant for the government's investigation of the explosion.

Importantly, on the periphery of all this possible explanation is a set of assumptions that remain fixed, including our beliefs that shuttles do not explode randomly, that events are caused by other events, and that this event happened in the

[51] According to Gleick's account (ibid.).

same space and time as all other events. Surrounding any given opportunity for explanation is an enormously complicated web of relationships, any particular thread of which we could follow to see that it too expresses logical relationships and patterns. We take this systematic and holistic context for granted when we ask why my plant is dead, why there is traffic, and why the *Challenger* exploded mid-flight.

Hence, our practices of reasoning, explaining, and inferring need to meet high requirements to get off the ground: all such activities presuppose a "totality of conditions" in Kant's language, or a holistic, systematic context. But Kant goes even further: he says that the inferential use of reason will inform us about "pure" reason (A 306/B 363)—which is the type of reasoning we do based on principles that have their source in "reason alone" as opposed to experience (A 305/B 362). As we will see, the very fact that we presuppose a holistic, systematically structured context in which our explanations take place is partly responsible for transcendental illusion—the fact that one naturally and inevitably assumes one has a soul, that there is a whole universe, and that God exists. It is precisely because these ideas arise naturally and inevitably out of our capacity to reason that we have the illusion that we can arrive at knowledge by reasoning "purely." First, we need to understand exactly how the ideas of the soul, the universe, and God arise from the nature of our reason.

4. Presupposing the Transcendental Ideas of Reason

To understand why the ideas of the soul, the universe as a whole, and God are necessary and inevitable we need to understand their role in our capacity to reason, which, as I have just argued, necessarily presupposes a holistic, systematic context. In

this section I argue that the transcendental ideas of reason are necessary and fundamental parts of that holistic and systematic context. This claim is based on something very simple on the surface: that there are three fundamental things necessary for the activity of explanation to get off the ground: a subject who explains, an object that gets explained, and a third thing that explains them both. These three things, I will claim, are just inchoate forms of the transcendental ideas of reason. When we attempt to *articulate* what it is that we presuppose, we end up with the ideas of the soul, the whole universe, and God. This shows that the transcendental ideas of reason are not elaborate philosophical constructions far-removed from everyday life, but rather necessary presuppositions, albeit incipient and unarticulated ones. It is only because we presuppose them that metaphysicians can set themselves the task of *knowing* what they are like.

Recall that the transcendental ideas of reason are peculiar types of ideas: peculiar because they attempt to represent the "totalities of our experience" (*P*, 4:328, see also A 327/B 383-4). I will argue here that by "totality," or "whole" [*Ganze*], Kant means that they attempt to represent the necessary components that go into thinking of our experience as a whole, complete entity. This means that all the components of experience would fit together not only systematically, but also in a way that yields a complete explanation.

Textual evidence that Kant thinks this way is given in several passages: "each individual experience," he says in the *Prolegomena*, "is only a part of the whole sphere of the domain of experience" (*P*, 4:328). The idea of a whole of experience is ultimately what metaphysics is attempting to capture—a point Kant puts by saying that

"metaphysics is a philosophy which is to present that cognition [all pure *a priori* cognition] in [a] systematic unity" (A 845/B 873). In other words, metaphysics, as a philosophical activity, attempts to grasp both the "parts" of experience and their relationship to the idea we have of the whole of experience. The idea of a whole of experience is simply the idea that everything that could possibly show up in experience shares systematic connections with other things, which is to say that all of these things can be explained by appeal to ever-more general principles.

This idea of a systematic whole is not only reflected in the systematic logical connections that individual experiences share, it also includes what we find essential for our moral and religious lives. The idea of a whole of experience would not be complete, for example, if it only captured mechanistic relationships among objects in space: it also needs to capture what Kant thinks is essential to our humanity. We need to see the world as the type of place in which our actions are meaningful, in which each person can be held responsible, and so on.[52]

We should not take this to mean that Kant thinks that the world itself *is* a complete, total, whole entity that allows total and complete explanations for everything in it. His point is that we have an *idea* that the world is whole and complete in this way—that our "individual experiences" share systematic connections to other individual experiences and that these are part of a larger systematic whole—an idea he thinks we presuppose when we ask for or offer an explanation of the *Challenger* explosion or why

[52] Kant insists that when experience is viewed this way, i.e., as the idea of a complete and whole sphere in which everything has its part and which includes everything necessary for our humanity, we can recognize if any crucial component is missing (A 832/B 860).

my plant is dead. To say that the transcendental ideas of reason attempt to represent the "totality of experience," then, is just to say that they are the most fundamental, necessary components of the *idea* of a whole, complete experience that allows for good and complete explanations of everything in it.

To begin to see this, we can start with the fact that experience needs a subject and an object (we will see how the idea of God comes in later). A passage from the *Opus Postumum* is particularly illuminating:

> Where does this scale of ideas come from? The *totality* of beings is a concept given *a priori* to reason, arising from the consciousness of myself. I must have objects of my thinking and apprehend them; otherwise, I am *unconscious* of myself […] for without that, I would be thoughtless, even with a given intuition, like an animal, without knowing that I am. (*Op* 21: 82)

Kant is saying here that in order to think at all, I must be able to be conscious of myself. Those familiar with the *Critique* will recognize this as a version of Kant's argument in the chapter called the "Transcendental Deduction" (B 129 ff). In order to think, the argument goes, I need to be capable of being conscious of myself. In order to be conscious of myself, I need to be conscious of objects; I could not differentiate *my*self if there is nothing from which I can distinguish myself—my thinking becomes *mine* in part because I can recognize my thoughts as separate from the objects I think about. Likewise, my thinking can be about objects in part because I can recognize my subjective thoughts as distinct from all my thoughts that are about real objects, i.e., my objective thoughts. Hence, in order to think at all, the thinker must have the most general idea of a subject and an object.[53]

[53] Further textual evidence supporting the claim that thinking requires the ideas of a subject and object is also given in Kant's unpublished "reflections" (the *Reflexionen*). There, he states that ideas, "as necessary

Of course, the most general idea of a subject and object do not yet amount to the ideas of the soul and the universe. A soul is not just a grammatical subject, for instance, the way that "table" is the subject of the statement "the table is brown." It is not just a substance either, insofar as substances are conceived as being subjects of predication, i.e., it is not just that a soul is a substance and not a predicate. Indeed, the soul is the subject of all possible thoughts, making the "I" a subject of a special type—what Kant calls the "complete" [*vollständigen*] subject (*P*, 4:330). Likewise the universe as a whole is a special type of object—it is not just an object the way say, an apple is. Rather, it is the object that contains within it, for lack of a better way of stating it, all other objects and events or states of affairs in their entirety. As Kant calls it, it is the "idea of the complete series of conditions" (ibid.).

We can begin to see now why the necessary distinction between a subject and an object naturally turns into ideas of the soul and the universe: in the context of thinking about what ultimately has to be the case in order for any one thing to be explained, we see that the "totality of conditions"—the systematic and holistic context we presuppose for our explanatory purposes—must relate somehow to the most fundamental explanatory conditions, which means all conditions must relate somehow to a subject and to a series of other objects. All "appearances" are going to relate to a possible explainer insofar as they are possibly represented by a thinker. Hence, a thinker is the "ultimate subject" in

---

concepts of reason … contain the *necessary conditions of the entire use of the understanding*" (R 18:228, my emphasis). This passage implies that the ideas of a subject and object are necessarily presupposed when we engage in philosophical inquiry about the very conditions of knowledge and thought.

the sense that the representation of any and all appearances can ultimately share a relationship with that thinker as his or her possible thought.

Furthermore, since the thinker is the very one explaining those appearances, it is *not one of those things*. A thinking subject must distinguish itself from all objects in order to grasp the contrast necessary to understand itself as having thoughts about *objects*. Of course, a thinker can think about itself—but insofar as it does this it does not think about itself as a subject but as an object. This contrast between the thinker and the objects of thought will ultimately lead to the idea that the soul is not something that can be predicated of anything else (i.e., that the soul is not a property of another object)—a claim, we shall see, that is representative of philosophy treating the soul as an object of knowledge. Remember too that explanations presuppose that each "individual experience" is related to every other in a systematic, i.e., logical, way. This will give us the idea that all components of experience are just part of a *whole* object—set apart from the subject who explains.

We see now why this subject turns into the idea of the soul in the context of the necessary conditions for our explanatory practices. If everything shares a possible relationship with the subject, then the subject must be thought of as absolute—that is, as not able to be predicated of anything else, and not dependent on anything material since it can have a thought of anything material and hence must think of itself as separate from that material thing.[54] This is how the idea of the soul emerges as one of an immaterial, ultimate subject. Because everything shares this possible relationship to a possible

[54] This is precisely why Kant says that we can only think of ourselves as appearances as not as "we are in ourselves" (B 153).

subject, Kant refers to this relationship as part of what is "*universal* in every relation that our representations can have" (A 333/B 390, my emphasis).

This also helps us understand a puzzling term that Kant uses in reference to the transcendental ideas: the "unconditioned" (A 323/B 379). "We can think the universal only by means of abstraction from all restricting conditions," Kant says in a reflection. "Abstraction from the determinations of the self makes the I seem unconditioned" (R, 18:228). In other words, we recognize that everything—considered as a possible thought—has a possible relationship with a thinker. Insofar as we are thinking of this subject merely as the thinker of possible thoughts, we can completely abstract away "all of its determinations" and think of it as capable of existing by itself without reference to any other conditions—hence the term, "unconditioned." Furthermore, to call it "unconditioned" is to imply that it needs no explanation for its existence, the way say, possible thoughts do.

Just as a possible relationship to a thinker is part of what is "universal in every relation," each component of experience can share in a possible relationship with every other component. Unsurprisingly, these possible relationships can again be expressed in the three logical ways that Kant thinks terms of a judgment can be related: a component of experience can be an attribute of another (e.g., brown being an attribute of a table), depend on another (e.g., the glass of water spilled because I accidentally knocked it), or simply exist at the same time as another (e.g., the couch is next to the chair). Because each component shares a possible logical relationship with every other, we form an idea

of a whole of things existing in space and time that allows for all of these relationships—an idea of the universe as a whole.[55]

We can now begin to see how the idea of God emerges. A hint is given in the very same passage in which Kant lays out how the ideas of the soul and the universe as a whole emerge:

> Now what is universal in every relation that our representations can have is 1) the relation to the subject, 2) the relation to objects, and indeed either as appearances, or as objects of thinking in general. If we combine this subdivision with the above division, then all the relation of representations of which we can make either a concept or an idea are of three sorts: 1) the relation to the subject, 2) to the manifold of the object in appearance, and 3) to all things in general. (A 333-34/B 390-91)

This passage reiterates why we are led to the ideas of the soul and the universe as a whole and introduces the idea of God as a third thing to which everything else can fundamentally relate. Kant puts this point in terms of a contrast between two types of objects: appearances or objects in general. "Appearances" are just the objects we experience—objects that are made possible for us *as objects* precisely because we can think of them as having certain spatiotemporal relationships to us and to other objects. On the other hand, "all objects of thought" includes not only possible objects of experience for us, but *all* possible objects and all the possible ways in which those objects could relate.

---

[55] If one thinks about what this means—the idea of the universe as a whole—one will soon realize why Kant thinks that there is something unsettling about it—namely, that our idea is either going to be "too small" or "too large" for the object it attempts to represent. First, we could think of the universe as a whole as a self-contained whole with spatial and temporal boundaries. Second, we could think of it as an object unbounded in space and time—"whole" in the sense that it includes everything (A 486/B 514). Both ways of thinking about the universe as a whole seem to capture what we mean by a "whole universe" and yet both fall short of capturing what we really want to represent. This is why Kant says there is something "antinomial" or contradictory about the very idea of the universe as a whole—meaning that there are two (incompatible) ways of understanding what it would mean for the universe to be a whole. This is why Kant's critique of the idea is called the "Antinomy" (A 405/B 432 ff.)

The contrast Kant is drawing here is tricky: anything humans can think of as an object in space and time is going to be an "appearance." We have only a negative idea of other possibilities, i.e., an idea that the way we experience objects is not the only possible way to experience them.[56] But this very contrast—the fact that there are other possible ways—other than our human way—to experience things, gives us the need for an explanation of why things are the way they are and not another way. This includes the explanation for the very possibility of a subject and an object—i.e., a soul and a universe. This is what Kant means when he says that the idea of God is the "thing that contains the supreme condition of the possibility of everything that can be thought (the being of all beings) (A 334/B 392).

The account I have given here helps motivate Kant's seemingly opaque claim that the transcendental ideas are connected to the three types of syllogisms. Kant recognized three types of syllogisms, one based on each of the relational functions: categorical, hypothetical, and disjunctive. Here I follow the account Kant gives in the *Prolegomena*, which is not exactly how he puts it in the first *Critique*. Kant says in the *Prolegomena*, "since I had found the origin of the categories in the four logical functions of judgments of the understanding, it was completely natural to look for the origin of the ideas in the three functions of syllogisms" (*P*, 4: 330). In the *Critique*, on the other hand, Kant connects each of the transcendental ideas with the forms of judgment as opposed to the forms of syllogisms (A 323/B 379). Even so, it is clear that even in the *Critique* Kant

[56] The contrast between appearances and objects in general is arguably the same contrast Kant makes between the appearance and the "thing-in-itself" or "phenomenon" and "noumenon," and is tightly connected with the contrast between a discursive intellect and a divine intellect (B 306-8).

wants to connect the transcendental ideas with *syllogisms.* He goes on to say, for instance, that each form of judgment leads to a type of syllogism (categorical, hypothetical, and disjunctive) and that in each of them "prosyllogisms proceed to the unconditioned" to each of the three transcendental ideas (ibid.). This implies that transcendental ideas are properly connected with syllogisms instead of merely with judgments.

Accordingly, we can expect that just as the concept of substance is connected with the logical function of a categorical judgment ("A is B") the idea of the soul will be connected with the function of a categorical syllogism, i.e., "All As are Bs, x is an A, therefore x is a B." Likewise, just as the concepts of cause and effect are connected with the logical function of a hypothetical judgment ("If A then B") the idea of the universe as a whole is supposedly connected with the function of a hypothetical syllogism, i.e., "If A then B, A, therefore, B." Furthermore, the idea of God is connected with the disjunctive syllogism, i.e., "Either A or B… not A, therefore, B."

Taken alone and out of the context of the more intuitive account I offered above, this account is admittedly puzzling. How is it, for example, that the idea of the soul is connected to a categorical syllogism? It is not obvious. At one point Kant implies that the transcendental ideas are each the result of a long string of prosyllogisms—as if reasoning about anything in the form of a categorical syllogism will eventually lead one to the idea of the soul (A 323/B 379-80). It is likewise not obvious that we get the idea of the whole universe from the logical form of judgment that expresses that one thing logically depends on another. Kant himself admits that the suggested connection

between the disjunctive syllogism and the idea of God will "at first glance" appear "extremely paradoxical" (A 336/B 393).

To illuminate this account, we should remember that the most promising way of understanding the connection of the transcendental ideas to our capacity to reason is that they are fundamental and necessary components of the idea of a whole and integrated experience. Above, I noted that in regards to the soul, because it is a thing that *everything* has a possible relationship with (in the way that a thought has a relationship with a thinker), we tend to think of this possible thinker in abstraction—i.e., not in relation to any particular thoughts that would specify any particular thinker. To repeat how Kant puts it: "we can think the universal only by means of abstraction from all restricting conditions" (NF, 18: 228). This is also true in regards to the ideas of the universe as a whole and God—since they are "universal" in the same way the soul is—i.e., they are among the things that *everything* can ultimately be related to—in order to think of them as "universal" in this sense, we must think of them abstractly—which is to say, as something general that could share in possible relationships but as something the thought of which does not share in *any particular* relationship. The transcendental ideas, in this sense, are abstractions—but, as we will see, abstractions of a particular type.

Kant's purpose in connecting the ideas with the forms of logical syllogisms is to show that the transcendental ideas are abstractions that pick up on the same logical patterns that emerge again and again in our thinking about the world. Just as the concept of substance expresses the logical relationship two things can share when one is predicated of another, the idea of the soul expresses that same relationship insofar as

thoughts are predicated of a thinker. The relationship that anything can have with a possible thinker can be logically expressed in a categorical form—i.e., a thought can belong to a thinker—just as an attribute belongs to a subject. Hence, we have an (admittedly loose) analogy with the categorical syllogism.

Likewise, the idea of the universe, as an abstract entity of a whole of things existing in space and time that allows for all the different types of relationships that the components of experience can share, loosely mirrors the hypothetical syllogism. This is because when we think of any given component of experience, we see that it necessarily *depends* on another—which is to say that "everything that happens (begins to be) presupposes something which it follows in accordance with a rule" (A 189/B 232)—or in other words, that every event has a cause. If we think about what depends on what—i.e., causes and effects, *through reason*, we come to the idea of a whole series of causes and effects, and hence an idea of a whole of space and time in which those causes and effects occur. This is again loosely connected with the hypothetical syllogism because this idea of the universe as a whole expresses logical relationships of dependence (i.e., If A then B, A, therefore B).[57]

The idea of God can be connected in the same way to the disjunctive syllogism. The idea of God, in this context, is the concept of the sum total of reality—the "highest being" (*ens summum*) (A 578/B 606) and the "most real being" (*ens realissimum*) (A

[57] Kant thinks the pure categories of the understanding are connected with the logical functions because it is the "same function that gives unity to the different representations in a judgment" that gives "unity to the mere synthesis of different representations in an intuition" (A 79/B 104-105). Kant uses the exact same language in regards to the ideas: the ideas could be found "nowhere else except *in this very act of reason* which, insofar as it related merely to form, constitutes the logical in syllogisms" (*P*, 4:330, my emphasis). Grier's account correctly emphasizes these passages, see (2001, 135).

576/B 604). The idea of an ultimate being that contains all of reality is what allows us our grasp of possibility in the same way ta disjunctive syllogism is an expression of our understanding that certain actualities cancel out certain possibilities. Our understanding that certain determinate qualities of an object rule out certain other determinate qualities, i.e., that a table is not both brown and not brown at the same time, is made possible through our grasp on what the possibilities are. God in this sense is the ground of all those possibilities.[58] This idea of God is not religious *per se*—Kant is not suggesting that we need to be theists in order to determine the qualities of any one particular thing. The idea is that in order to be able to assert that the folder in front of me is "red" as opposed to not red, I must have a grip on what *all* the possibilities *are*. The idea of God, insofar as it is construed as the most real being, is what allows me to have such a grip in the same logical way that I have a grip on reasoning about possibilities through a disjunctive syllogism.

We can now see why Kant connects each of the transcendental ideas of reason to a logical syllogism. We should not, however, put too much pressure on the analogy. Kant's main point, as I have shown, is that the soul, the universe as a whole, and God are necessary components of the idea of a holistic and systematic experience. His main point in connecting each transcendental idea with the form of a syllogism is to show that it is no coincidence that there are three such ideas and that they are what they are: for each idea expresses the same logical patterns that emerge in judgment.

[58] Again, Kant puts this in terms of activity: "*the act of reason* in disjunctive syllogisms is *the same in form* with that by which reason achieves the idea of a sum total of all reality" (*P*, 4: 330 fn, my emphasis).

Because the transcendental ideas of reason emerge necessarily for us as reasoners, we can see why Kant says that transcendental illusion—the beliefs that we exist as souls, that there is a whole universe out there, and that God exists—is necessary and inevitable. It is because we presuppose such ideas in our explanatory practices. It is precisely because the transcendental ideas are natural and inevitable that philosophers think they can gain knowledge of the soul, the universe as a whole, and God through reason alone, i.e., through *pure* reason.

5. Pure Reason: Legitimate and Illegitimate Uses

Kant compares transcendental illusion with the optical illusion that the moon is larger above the horizon than in mid-sky. Just as we cannot help but see the moon as larger above the horizon, we cannot help but believe that we exist as souls, that there is a whole universe out there, and that God exists. There is no getting rid of the illusion. We can, however, stop ourselves from making illegitimate judgments about the soul, the universe, and God, just as we can stop ourselves from mistaking the appearance of the moon as representative of its actual size. Transcendental illusion, as Kant says, cannot be "avoided at all, just as little… as the astronomer can prevent the rising moon from appearing larger to him, even when he is not deceived by this illusion" (A 297/B 354).

The necessity of the transcendental ideas of reason points to what Kant thinks is their genuine and legitimate use: to guide and regulate our organization of the components of experience in a way that allows them to fit together as a system and allows us to see them as part of an integrated whole. In other words, we can integrate novel experiences or discoveries into our belief system in the most optimal way if we act

as if thinkers are the absolute subject of thoughts, *as if* the universe had no beginning in space and time, and *as if* there is purposive design in the world (A 672/B 700). We seek explanations that fit together well with everything else we believe; mysteries are mysteries precisely because we do not yet have an explanation for them that fits consistently with everything else we believe. This is not to say that everything else we believe is infallible—it is to say that when we broaden our knowledge we adjust our beliefs in a way that attempts to fit all of our beliefs together in a systematic, holistic way.

Each transcendental idea, according to Kant, has a corresponding principle that tells us specifically how we ought to use it to guide the organization and unification of our knowledge. In Kant's language, each idea has a "schema" (A 674/B 702). The word "schema," as used in the Transcendental Analytic, is Kant's term for the type of representation that mediates between a pure concept of the understanding and a sensible representation; it is rule that allows us to apply a pure concept of the understanding to existing spatiotemporal objects (A 138/B 177). In the context of the Analytic, a schema is the product of the imagination as a faculty that mediates between sensibility and the understanding (A 142/B 181). Thus, when Kant calls the principles associated with the transcendental ideas of reason "schemas," he must mean something other than that they allow us to apply a pure concept of the understanding to an existing spatiotemporal object, since the ideas of reason do not represent spatiotemporal objects and because the pure concepts of the understanding are not applicable to them.

Indeed, to say that there is a schema associated with each transcendental idea of reason is to diverge from Kant's earlier use of the word, as he admits. He says, for example, that "it makes a big difference whether something is given to my reason as **an object absolutely** or is given only as an **object in the idea**" (A 670/B 698). In the second case, we have *only* a schema "for which no object is given… but which serves only to represent other objects to us, in accordance with their systematic unity, by means of the relation to this idea" (ibid.). What he means here is that the schema for each transcendental idea of reason does not itself represent an object. Rather, it is a rule that allows us to represent other objects in a systematically unified way. The schemata of the ideas then, are not products of the imagination. Presumably, they are products of *reason* insofar as they show us how to unify *judgments*, which themselves are products of the understanding.

Kant says that the psychological idea—the idea of the soul—allows us to "connect all appearances, actions, and receptivity of the mind to the guiding thread of inner experience **as if** the mind were a simple substance that (at least in this life) persists in existence with personal identity, while its states—to which the states of the body belong only as external conditions—are continuously changing" (A 672/B 700, Kant's emphasis). Acting as if the soul were a certain type of thing is what allows us to unify our knowledge.

It is clear that the idea of a holistic, integrated experience includes our moral commitments—so the transcendental ideas will not only regulate our organization of theoretical knowledge, it will regulate how that organization fits in with our moral and

religious commitments, since an idea of a whole experience would not be complete, according to Kant, without including those things. Recall that Kant thinks we have particular interests in virtue of being human. "Speculative" interests compel us to want good complete explanations for everything and "practical" interests compel us to see the world as a place in which our actions are efficacious and meaningful (see chapter one, section two).[59] Kant is clear that it is our human interests that ultimately guide how we organize our knowledge.

The idea of the soul then allows us not only to unify our speculative knowledge in the most optimal way, but also to treat ourselves as if we had identity across time so as to take responsibility for our actions, thereby serving one of our practical interests. Similarly, the ideas of the universe as a whole and of God guide the synthesis and unification of our knowledge in ways that serve both our speculative and practical interests. For example, part of our operative idea of the universe as a whole, according to Kant, is that it allows for both a free and determined cause. This idea regulates the unification and synthesis of our judgments by suggesting that we ought to unify and synthesize our judgments in a way that acknowledges and maintains the distinction between the two types of causes. For example, when reasoning about why a criminal engaged in destructive behavior, our story about his violent and sorry upbringing as a determinant factor in his character should not rule out that he was free and responsible. Unifying our judgments in such a way serves both our speculative and practical interests.

[59] As I said in chapter one, and as I will discuss more in chapter four, because our speculative interests ultimately lead to a contradiction when thinking about the nature of the universe, Kant thinks that our speculative interests can be subordinated to our practical interests, which amounts to saying, for Kant, that practical reason has primacy over speculative reason (*C2*, 5: 119 ff.).

This is what Kant means when he says that the ideas of reason have an indispensible "regulative use" (A 644/B 672).

I am now in a position to make more precise a claim from chapter one: that Kant's critical method reorients us in regards to our metaphysical questions. For Kant, the entire aim of metaphysics is to help us see clearly beyond our immediate experience and inform us about our vocation as human beings (B xix-xx, A 840/B 868). Taking the transcendental ideas to have a merely regulative use allows us to use them in a way that restores to philosophy its essential task: to illuminate that "which necessarily interests everyone" (A 839/B 867), the "essential ends of human reason" (A 839/B 867).

Philosophy can succeed in this task, however, only if we are not fooled by transcendental illusion into thinking that we can have knowledge that we cannot have. The regulative use of the transcendental ideas—acting *as if* we and the universe are certain types of things—does not amount to knowledge that the soul and the universe *are* those things. But as I have shown, the beliefs that one has a soul, that there is a universe as a whole, and that God exists, are necessary, inevitable, and unavoidable. The power of transcendental illusion tempts us to make knowledge claims that step beyond what is verifiable in experience—it tempts us to treat the soul, the universe, and God as objects that have knowable natures and properties—a temptation particularly strong for professional philosophers who aim to articulate the conditions of experience. Now, we might wonder why Kant does not just ignore such arguments or dismiss them as unsound and unworthy of analysis. Kant does not do this for a very important reason: being deceived by the transcendental illusion that we can have knowledge of the soul, the

universe, and God leads to arguments that not only are unsound but also stymie the aim of metaphysics. These arguments prevent us from understanding the legitimate and necessary use of these ideas. The next chapter discusses how transcendental illusion manages to deceive certain metaphysicians about the soul.

THE ILLUSION OF AN IMMATERIAL SELF: TRANSCENDENTAL REFLECTION AND THE SOUL

CHAPTER 3

> *The light dove, in free flight cutting through the air the resistance of which it feels, could get the idea that it could do even better in airless space. Likewise, Plato abandoned the world of the senses because it posed so many hindrances for the understanding, and dared to go beyond it on the wings of the ideas, in the empty space of pure understanding.*
>
> —Kant, *Critique of Pure Reason*, A 5/B 8

"Transcendental illusion," as I argued in the last chapter, is the illusion that there exists a soul, a whole universe, and a God. I argued that this illusion is deeply rooted in our explanatory practices and that it is therefore necessary and unavoidable. We only err, Kant thinks, when we take the transcendental ideas of reason to be significant not because our explanatory needs require us to, but because we take them to represent real, knowable objects. This chapter addresses how Kant thinks transcendental illusion deceives the "rational psychologists," or those philosophers who attempt to know what is true of a thinker from thought alone. In a chapter called the Paralogisms, Kant presents and critiques four arguments concerning the thinking subject, or "soul." He claims each argument is a fallacy.[60]

On a general level, commentators applaud Kant's critique of rational psychology. Strawson, for instance, states that it is philosophical criticism of the highest order.[61] While readers generally agree, however, that Kant's critique successfully demolishes the basic tenants of rational psychology, most complain that the details are unclear or

[60] I will refer to the chapter with the upper-case "Paralogisms" and the arguments themselves with the lower-case "paralogisms."

[61] Strawson 1975, 162.

confused. There is little agreement over what form the rational psychologist's arguments supposedly have or exactly what fallacy they commit. It remains unclear on all accounts, I will argue, what the exact connection is between the arguments of rational psychology and transcendental illusion.

In this chapter I argue that the paralogisms and Kant's critique of them are greatly illuminated if we rely on a resource given outside the Transcendental Dialectic, namely, Kant's theory of transcendental reflection that he develops in a chapter called the "Amphiboly." There, if we bother to recognize it, Kant gives us the resources to understand how the arguments of rational psychology arise, what motivates them, exactly how he thinks they go wrong, and how they are connected to transcendental illusion. While Kant does not specifically connect the paralogisms with his theory in the Amphiboly, I will show that the two must be connected and indeed, that doing so leads to a deeply coherent picture of how pure reason goes wrong in regards to the soul. Furthermore, understanding the paralogisms through the lens of the Amphiboly reveals how Kant's own method of critique guards us against being deceived by transcendental illusion, since it presents a method that allows us to identify the true significance of our judgments.

I begin in section 1 by summarizing the paralogisms and by showing why interpreters disagree about their form. In section 2 I introduce what I argue is the key to understanding the paralogisms, namely, a method Kant calls "transcendental reflection." I begin the discussion by summarizing the activity of "comparison," which Kant claims happens prior to making objective judgments. Comparison takes place via pairs of

concepts that Kant calls the concepts of comparison. I summarize these concepts in section 2 and show how Kant thinks they ground objective judgments. The problem is that these concepts can be used to compare representations in one of two ways, which means that the judgments they ground have two possible meanings. I discuss this two-fold use of the concepts of comparison in section 3. This will allow us to see how transcendental reflection helps one to recognize how one's pre-judgmental comparisons inform objective judgments, which is what I discuss in section 4. As I will continue to argue in section 5, the failure of the rational psychologists to transcendentally reflect is precisely why they are fooled by the transcendental illusion that the soul is an object. I proceed by showing how the lack of transcendental reflection grounds the fallacies of each of the first three paralogisms. In section 6, I show how it applies to the fourth paralogism. I end in section 7 with a discussion of what this all means for Kant's positive theory of the self.

1. The Paralogisms of Pure Reason: The Arguments and Their Various Interpretations

Rational psychology is a discipline that claims to know what the soul is like merely from the phrase "I think." Descartes, perhaps the most notable example of such a philosopher, claims in the *Meditations* that we can know that the soul is a simple, indivisible substance "really distinct" from the body.[62] In the Paralogisms Kant addresses four claims included in Descartes' position: that the soul is a substance, that it is simple,

[62] To say that the soul is "really distinct" from the body is to say that the two can exist separately (Descartes 1998, 115).

that it is identical over time, and that it is independent from all physical objects including the body. These arguments comprise the four paralogisms respectively.

Substance-hood, simplicity, personality, and independence, are all building-blocks, Kant thinks, for more meaningful conclusions about the soul. The rational psychologist thinks the soul's substantiality implies its immateriality. Likewise, the soul's simplicity and substantiality together imply its incorruptibility and, along with personality, its spirituality and immortality (A 345/B 403). Kant identifies simplicity as the "Achilles of all dialectical inferences" regarding the soul—a claim he supports in his logic lectures where he states that simplicity implies indivisibility, which implies imperishability, which implies persistence (*The Hechsel Logic*, 112).[63]

Kant presents each argument as a syllogism. A syllogism is an argument with two premises, the "major" and "minor," and a conclusion. The first three paralogisms are all categorical syllogisms, which means the major premise is of the form "All *F*s are *G*s," the minor premise of the form "*a* is an *F*," which together entail a conclusion of the form "*a* is a *G*." A syllogism has three "terms," *F*, *G*, and *a*, the middle, the major, and the minor respectively. In each of the first three paralogisms, the minor term is the "I" of "I think." The conclusion then predicates something of this "I," namely, its substantiality, simplicity, and personality, respectively.[64] For example, the first paralogism states:

> [1] That the representation of which is the **absolute subject** of our judgments, and hence cannot be used as the determination of another thing, is **substance**.

[63] The simplicity of the soul will take center stage in the next chapter, where I discuss in more detail Kant's critiques of arguments for and against it.

[64] Since the fourth paralogism has a slightly different form, I delay discussing it until section 6.

> [2] I, as a thinking being, am the **absolute subject** of all my possible judgments, and this representation of Myself cannot be used as the predicate of another thing.
> Thus I, as a thinking being (soul), am **substance.** (A 348, Kant's emphases)

The major premise contains the major term—substance—and the middle term—"that the representation of which is the absolute subject of our judgments and hence cannot be used as the determination of another thing." The minor premise then predicates the middle term of the "I" of "I think." This entails the conclusion that "I, as a thinking being (soul) am substance."

The second paralogism states:

> That thing whose action can never be regarded as the concurrence of many acting things, is simple.
> Now the soul, or the thinking I, is such a thing.
> Thus etc. (A 351)

Although Kant does not bother finishing it, presumably the major term is "simple," the middle term is "that thing whose action can never be regarded as the concurrence of many acting things," and the minor term is "the soul or the thinking I." The implied conclusion is that the thinking I (or soul) is simple. Likewise, the third paralogism asserts that the soul is a person through the middle term "what is conscious of the numerical identity of its Self in different times" (A 361).

Kant believes these arguments are fallacies. But his account of exactly where they go wrong has troubled commentators. He identifies a paralogism as a formal fallacy, which supposedly means that it has an invalid form (A 341/B 399). But straight-forwardly, the arguments are valid. Furthermore, Kant says a paralogism commits the fallacy of "*sophisma figurae dictionis*" (A 402-3/B 411), which he identifies elsewhere as

the fallacy of the ambiguous middle (*The Hechsel Logic*, 110-11).[65] This indicates that the middle term of each paralogism is ambiguous. But he goes on to imply that the *major* term (substance, simplicity, and personality respectively) is ambiguous, insofar as it is being used "transcendentally" in the major premise and "empirically" in the minor (A 402-3). To complicate things further, his account of the fallacy seems to change in the second edition. There, he says that while the major premise is about a "being that can be thought of in every respect" the minor one is only about this being "insofar as it is considered as subject." Additionally, he states that it is not the major term that is ambiguous but rather the term "thinking" (B 411fn), which is not precisely *any* of the terms, although we might think he is referring imprecisely to the minor. Thus, we have two related issues in regards to Kant's analysis of the forms of the paralogisms: 1) whether they are valid or invalid, and hence whether they commit a formal or informal fallacy, and 2) if they indeed commit the fallacy of the ambiguous middle, which of the terms is ambiguous.

In addition to the confusion about the form of the arguments, Kant's evaluation of their content is also unclear. He states that the conclusion that the soul is a substance is "valid," so long as we take ourselves to be a substance "only in the idea but not in reality" (A 350-51). He also states that we can "pretend to know that the thinking I … is simple" so long as we do not understand that claim to "extend our cognition in the least" (A 361). Similarly, he says that the concept of "personality" can "remain" so long as we never "boast of it as an extension of our self-knowledge through pure reason" (A 365-

[65] Kant classifies a fallacy of the ambiguous middle as a formal fallacy (*Dohna-Wundlacken Logic*, 777).

66). These comments indicate that Kant has no problem with the belief that the soul is *in some sense* a simple substance identical over time. It is unclear, however, which claims of the arguments he thereby endorses. It could be that Kant agrees with the conclusions of each argument and that his critique merely shows that they ought not to be inflated. Or it could be that Kant agrees with the minor premises of each argument and that his critique shows that we can only understand it to be a merely logical claim and not a "transcendental" one. Hence, in addition to the two issues regarding the form of the paralogisms, we have a third, which concerns the content, namely, the issue of which of the claims of rational psychology are unproblematic.

Kant's imprecise evaluation leads to very different interpretations, all of which are supported by the text. We can see why several commentators, on the one hand, claim that the arguments themselves are valid and hence do not commit a formal fallacy.[66] We can see on the other hand, why some insist that we must maintain an interpretation that at least understands the paralogisms as committing the fallacy of the ambiguous middle (Proops 2010, 470). Likewise, we can see why many commentators think that because the arguments are valid, it must be their content that Kant finds at fault. Bennett, for example, argues that Kant agrees with the content of the premises but that the fault lies in its inflated conclusion (1967, 72).[67] Some commentators believe that the fault lies in the *first* premise, where the category (substance, simplicity, or personality) is used to

[66] See Van Cleve 2012, 483; Ameriks 1982, 67; and Bennett 1967, 72. Hughes, for one, implies that only a "logical" paralogism is formally invalid and that a transcendental one is not (1983, 405).

[67] Ameriks offers a similar way of understanding them, although in the context of presenting a variety of possible ways of understanding the fallacies (1982, 56).

represent a claim taken mistakenly as a synthetic *a priori* truth, while the second premise is merely a tautological claim about the logical subject of thought that Kant himself endorses.[68] On the other hand, Kitcher argues that the major premise is merely a definition that Kant himself agrees with and that some version of the second premise is also agreeable to Kant (1982, 519-20).[69] Grier argues that two types of ambiguity are operating, one that is characterized by the formal fallacy of the ambiguous middle and another that is rooted in transcendental illusion and manifests itself in the second premise (2001, 157-58).

Quite apart from these formal and informal issues surrounding the paralogisms, another issue is their exact connection to transcendental illusion. I have mentioned that before Grier's account, commentators tended to elide transcendental illusion with the arguments themselves (for her own discussion, see Grier 2001, 9-10). Grier correctly argues that doing so is a mistake: Kant clearly says that transcendental illusion is unavoidable and necessary, while the fallacies of rational psychology are avoidable (ibid., 149). As such, most commentators have now adopted Grier's position that transcendental illusion and the dialectical inferences are separate and hence independent of each other. But this independence reveals a gap that needs explanation. If the fallacious arguments and transcendental illusion are independent, what exactly is their connection?

Grier's answer is that the fallacies, while grounded in transcendental illusion, are "independently motivated by a transcendental misapplication of the categories" that

[68] Allison 2004, 284; Buroker, 2006, 215-16.

[69] Similarly, Proops argues that "any adequate reading must portray the paralogism's premises as true" (2010, 470).

issues from a "failure to draw the transcendental distinction between appearances and things in themselves" (ibid., 153). Recall that for Kant a "transcendental realist" is one who believes that appearances are things in themselves. Hence, a rational psychologist is a type of transcendental realist. On this interpretation then, the mistake of the rational psychologist is to fail to recognize that the "I" of "I think" represents only an appearance and not a thing in itself. This mistake, according to Grier, *in combination* with transcendental illusion, is what leads to the paralogisms. For Grier, transcendental illusion causes the second premise of each paralogism to be ambiguous, which in turn "provides the basis for the paralogistic fallacy" (ibid., 159). So Grier's account is that the rational psychologists are first misled by transcendental illusion and then, because they mistake the thing in itself for the appearance, make faulty inferences about the soul.

But we should take a closer look at how this mistake supposedly happens. On Grier's account transcendental illusion causes the second premise of each paralogism to be ambiguous. As she says, the minor premise is understood in one sense to be about the "representation 'I'" insofar as it is understood as asserting the "necessary possibility of attaching the 'I' of the 'I think' to all my judgments." Insofar as this is true, the minor premise "simply reiterates the principle of apperception found in the first part of the B Deduction" (ibid., 158). In another sense, however, the minor premise is understood as representing a metaphysical self or soul. This is a promising way of interpreting the second premise. But we should recall that on Grier's own account, this confusion, the taking of the "transcendental representation of the self" as a "representation of a metaphysical self," is just what transcendental illusion amounts to, which means that this

very confusion is what is supposed to be natural and unavoidable. On this view, the rational psychologists only make a *mistake* when they "determine" this "pseudo-object" through the category of substance (ibid., 159). In other words, the mistake of rational psychology is to think that it can reason about something it takes to be a metaphysical object.

It seems to me that this gets things backwards. To identify the fault as one of *reasoning* about what is in actuality a 'pseudo-object' as if one can gain knowledge of that object surely misidentifies the location of the mistake. The rationalist psychologist does not go wrong just because they think they can *reason* about something they take to be an object. Rather, they go wrong in *taking something* to be an object that is not an object at all. The mistake is to fail to recognize that the object they are reasoning about is indeed only a "pseudo-object" to begin with. On Grier's account, however, taking the second premise as the representation of a metaphysical object instead of as a logical truth about the "I think," just *is* transcendental illusion. So her account seems to imply that the illusion *itself* leads to the faulty inferences, not, as she claims, the failure to distinguish between the thing in itself and the appearance.

Another way of making the same point is to realize that on Grier's account it is unclear how the transcendental idealist—the one who *does* distinguish between the thing in itself and the appearance—can avoid making the same faulty inferences as the transcendental realist. For transcendental illusion, on Kant's account (and on Grier's) is necessary and unavoidable. Thus, if transcendental illusion is responsible for the ambiguity of the second premise, there seems to be nothing that would stop even the

transcendental idealist from reasoning about the second premise in the same way that the rational psychologist does. This is because *there is nothing in transcendental idealism that says that one cannot reason about or gain knowledge of objects*. If a transcendental idealist takes the soul to be an object, which is what transcendental illusion gives him the "propensity"[70] to do according to Grier, further distinguishing between things in themselves and appearances will not stop him from making faulty inferences; the distinction would simply come too late. So on Grier's reading, it is unclear exactly how a transcendental idealist avoids being deceived by transcendental illusion.

I am not claiming here that we need an account whereby the transcendental idealist should be able to avoid transcendental illusion. On Kant's account, transcendental illusion—the illusion that there exists a soul, a whole cosmos, and a God—is unavoidable regardless of whether or not one makes a distinction between the thing in itself and the appearance. But we do want an account of the paralogisms and transcendental illusion that allows the transcendental idealist to avoid being *deceived* by the illusion. The transcendental idealist should be in a position to see transcendental illusion as illusory.

It appears that even on Grier's groundbreaking account, we are left wondering what the connections are between 1) the faulty arguments of rational psychology, 2) transcendental illusion, 3) transcendental realism, and 4) transcendental idealism. The account I offer in this chapter not only allows us to draw a robust connection between

[70] Grier thinks transcendental illusion is ultimately a "propensity to take the subjective or logical requirement that there be a complete unity of thought to be a requirement to which 'objects' considered independently of the conditions of experience (things in themselves) must conform (A 297/B 354)" (2001, 8).

these components of Kant's account, it also allows us to see precisely how transcendental idealism gives us the resources to recognize transcendental illusion *as such*, while simultaneously allowing for the necessity and inevitability of that illusion.

On my reading we will see that it is indeed premise two that is the locus of the mistake and that therefore, we can maintain some of the insights of Grier's reading. We will see, however, that it is not transcendental illusion that drives the ambiguity of the second premise, as Grier implies, but rather the assumptions of transcendental realism. The transcendental realists do not have the resources in their ontology to recognize that the second premise is ambiguous and, *because of this*, are deceived by transcendental illusion. So, while the transcendental idealists still find it necessary to think of the soul as an object, they are not misled into thinking they can gain knowledge of it as such, since they are in the position to acknowledge that the transcendental idea of the soul is significant only because of its requirement for our explanatory practices. The key is to be found in Kant's theory of "transcendental reflection."

2. Introduction to Transcendental Reflection: the Concepts of Comparison

In this section, I will introduce Kant's theory of transcendental reflection, which I argue is the key to understanding how the paralogisms arise for the rational psychologist. Kant's theory requires some unpacking and it is only over the next few sections that we will see how the activity of transcendental reflection can help us in regards to understanding Kant's complaint against rational psychology. We can begin by interpreting Kant's initial remarks on transcendental reflection. To begin, he says that transcendental reflection is:

> the action through which I make the comparison of representations in general with the cognitive power in which they are situated, and through which I distinguish whether they are to be compared to one another as belonging to the pure understanding or to pure intuition. (A 261/B 317)

First, Kant references a "comparison of representations." What does this mean? He goes on to say that "prior to all objective judgments we compare the concepts" (A 261/B 317). So a comparison of representations is an activity that happens *before* we make an objective judgment. We use certain concepts to engage in such comparisons. He gives us eight such concepts, presented in pairs: identity and difference, agreement and opposition, inner and outer, and the determinable and the determination (or matter and form). He calls these concepts the "concepts of comparison" (A 262/B 318).

In the next section, I offer an interpretation of exactly how these concepts are used to compare representations. First, though, we should get the big picture of how such comparisons operate prior to judgments. Kant claims that the way in which we compare two representations will *determine* what form a judgment will take:

> with regard to **identity** (of many representations under one concept) for the sake of **universal** judgments, or their **difference**, for the generation of **particular** ones, with regard to **agreement**, for **affirmative** judgments, or **opposition**, for negative ones, etc. (ibid)

Here, Kant aligns the concepts of comparison with the logical form of judgments. Earlier in the *Critique*, Kant offered a table called the "logical functions of judgment" (A 70/B 95). The table is divided into four broad classifications, each capturing a possible logical aspect of a judgment: quantity, quality, relation, and modality.[71] The "quantity" involves the number of things the judgment is about, the "quality" involves whether the judgment

[71] Just as with the paralogisms, the fourth category here, "modality," is different from the rest. I delay discussing it until section 6 where I discuss the fourth paralogism.

affirms something or denies it, and the "relation" involves what kind of relationship is asserted between the terms of the judgment. Under each of these headings Kant offers three "functions." The "quantity" of a judgment can be "universal," meaning it predicates something of "all" of a certain thing, "particular" meaning it predicates something of "some" of a thing, or "singular," which predicates something of one individual thing. Likewise, the "quality" of a judgment can be "affirmative," meaning it affirms something ("it is the case that…"), "negative," meaning it denies something ("it is not the case…"), or "infinite," meaning it neither affirms nor denies (*F*s are non-*G*s). Similarly, the "relational" aspect of a judgment can be "categorical" (*F*s are *G*s), "hypothetical," (If *p* then *q*), or disjunctive ("*F* or *G* or *H*….").

We can see in the above passage that Kant deliberately aligns the concepts of comparison with the logical forms of judgment. He states that prior to making a judgment of "universal" form, for example, we compare the concepts in regard to their "identity." Likewise, a comparison of concepts in regards to their "difference" takes place prior to a judgment of particular form.[72] Similarly, the concepts of "agreement" and "opposition" are respectively aligned with the logical forms of affirmation and negation.

As such, a judgment of the form "All *A*s are *B*s" is aligned with the concepts of "identity" and "agreement," since it is a universal affirmative judgment. Likewise, a judgment of the form "Some *A*s are not *B*s" would be aligned with the concepts of

[72] We might ask why there is no concept of reflection aligned with the "singular" form. I will provide some insight about the omission below. As Longuenesse points out, for Kant, the third listing under each heading does not properly represent "distinct logical forms. For from a strictly logical standpoint, the singular is not distinguished from the universal judgment because in both alike the predicate is attributed to the totality of what is thought under the subject" (1998, 139).

"difference" and "opposition," since it is a particular negative judgment. This is all Kant states directly. Frustratingly, he does not finish the sentence in which he associates the forms of judgment and concepts of reflection, leaving it to the reader to interpret what he means by the word "etc." So, while we can be sure that Kant means to align the concepts of identity and difference with the quantity aspect of a judgment and the concepts of agreement and opposition with the quality aspect of a judgment, we are less sure that he intends for us to go on in a similar fashion with the rest of the concepts and logical forms. Here, I follow Longuenesse's argument that Kant does in fact intend us to proceed that way.[73]

Assuming that Kant means to align the rest of the concepts of reflection with the rest of the logical functions, we can infer how those connections would go. The concepts of inner and outer would be aligned with the relational aspects of the categorical and hypothetical forms respectively. A categorical judgment of the form "*A* is *B*" would be aligned with the concept of inner. Likewise, a hypothetical judgment of the form "if *A* then *B*" would be aligned with the concept of outer. Here, we should note that a hypothetical judgment does not seem to have a quality or quantity, so does not manifest *all* the aspects of a judgment. Similarly, we should not expect it to align with exactly four concepts of reflection. At any rate, it seems as though a judgment will be aligned with as many concepts of reflection as it is with aspects of a judgment. To finish the

[73] An opposing view comes from Klaus Reich, who argues that Kant meant to align concepts of comparison only with the quantity and quality aspects of a judgment and not relation and modality (Reich 1992, 80-82). But as Longuenesse makes clear, Kant repeatedly asserts that "there are as many concepts of comparison as there are logical forms of judgments" (Longuenesse 1998, 128-29).

association, we can note that the concepts of determinable and determination (or matter and form) would align with the modal aspects of possibility and actuality.

We must get more definite about what it means for these concepts to guide us in a comparison of representations "prior to all objective judgments." The suggestion seems to be that before making an objective judgment, we engage in a comparison of concepts, presumably those that we are going to relate in a judgment. Thus, a judgment like "all rectangles are squares," arises first by *comparing* rectangles and squares, via some of the concepts of comparison. Because this is a categorical, universal, and affirmative judgment, the concepts of comparison that will presumably be involved are *inner*, *identity*, and *agreement*. Kant's claim seems to be that we use the concepts of reflection to judge how our two concepts (rectangles and squares) relate, in order to yield a judgment.

We are now in a better position to understand what Kant means when he says that transcendental reflection is the "action through which … I distinguish whether they [representations] are to be compared to one another as belonging to the pure understanding or to pure intuition" (A 261/B 317). I transcendentally reflect when I discern *how* I compare two representations. Here, Kant does *not* mean that I need to discern how two representations relate in the form of a judgment. That is, he is not suggesting that I track how the concepts of "cat" and "black" are related in the judgment that some cats are black. *That* relationship is captured in the logical functions of judgment that Kant labels under the term "relation." Instead, the suggestion is that I ought to track how the representations are being compared *pre*-judgment. Transcendental

reflection is to recognize whether the two representations I compare belong to one another in the "pure understanding" or the "pure intuition." So our next task is to figure out what that means.

Kant goes on to clarify that he is speaking of the pure understanding or pure intuition as faculties or "cognitive powers" (ibid.). The faculty of understanding, of course, refers to our capacity to use concepts, while the faculty of intuition refers to our sensible capacity. The "pure" understanding would then refer solely to conceptual representations untainted by anything spatiotemporal, while the "pure" intuition would refer solely to our spatiotemporal representations, untainted by any conceptual representations. So Kant is saying that transcendental reflection is an activity whereby we discern whether we compare two representations in a merely conceptual way or in a merely spatiotemporally way.

Of course, we now want to know what it means to consider objects merely conceptually or merely spatiotemporally. We also need to know why it is important to track precisely how one compares representations prior to a judgment. The answer to the first question will come in section 4 below. The short answer to the last question is that we need to track it because the concepts we use to compare two representations lead to ambiguous judgments otherwise—ambiguous insofar as they do not by themselves express *how* they are true, whether of representations considered merely conceptually or of representations considered merely spatiotemporally. *How* we compare two representations, Kant is indicating, will make a difference in discerning the truth of the judgment; in particular, it will allow us to correctly identify which judgments are proper

synthetic *a priori* judgments and those that are not but might appear to be (Kant makes this point at A 263/B 319). In the next section, I explicate in detail just how each pair of concepts expresses the logical functions of judgment. Then, we will be able to see, in section 4, how each of these pairs grounds potentially misleading judgments.

3. The Concepts of Comparison and The Logical Forms of Judgment

The alignment between the concepts of comparison and the logical functions of judgment is as follows:

QUANTITY

Logical forms of judgment: *Universal/Particular*
Concepts of reflection: *Identity/Difference*

QUALITY	**RELATION**
Logical forms of judgment: *Affirmative/Negative*	Logical forms of judgment: *Categorical/Hypothetical*
Concepts of reflection: *Agreement/Opposition*	Concepts of reflection: *Inner/Outer*

MODALITY

Logical forms of judgment: *Problematic/Assertoric*
Concepts of reflection: *Matter/Form*

Now the task is to understand just how each pair of the concepts of reflection is used to ground the different possible aspects of a judgment. Kant's idea seems to be that we compare two representations in one or more ways in order to yield an objective judgment (i.e., a judgment that has one or more of the possible logical forms presented in the logical table of judgments). The way that the concepts of comparison are aligned with the logical aspects of a judgment implies that we can compare those representations in one of four ways (perhaps all four): in regards to their quantity, quality, relation, and

modality. Here, I will offer details of how these logical forms of judgment are grounded in comparisons via the concepts of comparison. I analyze each aspect of a judgment independently with the crucial caveat that any such treatment is doomed to be artificial. All judgments depend on several comparisons that inevitably depend on one another and overlap. But with this in mind, we can get an idea of how these pre-judgment comparisons operate. For various reasons, I begin with the concepts of inner and outer.

a. Comparing Representations According to Their Relation: The Concepts of Inner and Outer

One of the most basic comparisons we can make prior to making a judgment is by comparing how representations relate *to each other* through the concepts of inner and outer. This comparison simply involves whether one is inner or outer to the other. Kant himself is unclear about what this means on any precise level, but I propose here that two representations would share the relationship of "inner" if one is possibly attributable to another as a predicate, without any conditions added. Likewise, they would share the relationships of "outer" if one is possibly attributable to the other as a predicate, but only under stated conditions.[74]

[74] I use the word "possibly" here because as we will see, this type of comparison does not yet yield an affirmation or denial in any way. It is the "quality" aspect of a judgment that does that. My interpretation of inner and outer here is similar to that of Longuenesse, who argues that to so compare is to "attend to the character of the discursive condition (the condition expressed in a concept or combination of concepts) under which a predicate can be attributed to a subject. In other words, it is to consider whether a predicate can be attributed *nulla adjecta conditione* (no condition being added) or only *sub adjecta conditione* (under an added condition)" (1998, 141-42). I deliberately leave my interpretation more open than that of Longuenesse's, however. While she argues that this type of comparison is to see if a condition is expressed in a *concept*, I will show below that we must interpret Kant more broadly here. This is because there will be two ways of comparing whether or not a representation is "inner" or "outer" to another—a conceptual way *and* a spatiotemporal way. The latter involves comparing whether one "representation," i.e., property is literally outer or inner to another (where the second representation is of an *object*).

Evidence for my interpretation is given in remarks that Kant makes in his logic lectures about the concepts of inner and outer. There, he says that an inner determination means something like a "mode" or an accident—a property that is "extra-essential" because it is not considered part of the essential concept of a thing. As he says there, "*extra-essential* marks are again of *two kinds*; they concern either *internal* determinations of a thing (*modi*) or its external relations (*relationes*). Thus the mark of *learnedness* signifies an inner determination of man, but *being a master or a servant* only an external relation" (*The Vienna Logic*, 839). Here, he makes clear that the concept of inner captures what is determinative of the subject without any conditions stated, and the concept of outer captures what is determinative of the subject only via reference to something outside of it, i.e., a condition. So a relational comparison would be one that determines whether or not one representation is related to the other as a mode (an internal determination of that thing) or relation.[75]

The implication is that if the comparison shows that one representation is "inner" to the other, i.e., a mode or internal determination, one will make a judgment that has a categorical form, i.e., "*A* is (or is not) *B*." Likewise, if the comparison shows that one representation is "outer" to the other, i.e., that they only share an external relationship,

[75] We might ask here how we know which representation determines the other. This question brings out the artificiality of any systematic presentation of such concepts. For it is precisely *another* act of comparison that determines which representation is the one that gets determined by the other, namely, the concepts of determinable and determination (i.e., form and matter). For various reasons, however, it would have simply been too unwieldy to begin our discussion with those concepts.

one will make a judgment that has a hypothetical form, i.e., "If *A* then *B*" or "It is not the case that 'If *A* then *B*'."[76]

This can be made clearer by discussing examples. The interpretation I have offered indicates that for Kant, a judgment such as *the desk is brown* has presupposed a comparison that showed that brownness is possibly attributable to the desk with no conditions added, while a judgment such as *if it rains the pavement gets wet* has presupposed a comparison that showed that wetness and the pavement share a possible external relationship. In other words, wetness is not attributable to the pavement unless another condition is met, namely, if it rains. The property of wetness is thus external to the pavement. It does not belong to the pavement as an inner determination or accident.

We might pause here to ask whether or not we could just as well say "the pavement is wet," turning such a judgment into a categorical one instead of a hypothetical one. A judgment like the "pavement is wet" would presuppose a comparison that wetness is an inner determination or mode of the pavement, as opposed to a comparison that wetness is an external determination that depends on the condition of rain. The question would be how we possibly determine whether or not we understand wetness as an internal mode or as a property possible under external conditions. For surely, we could also say that the "if it was made of wood, then the table is brown." This difficulty again highlights the artificiality of treating each logical aspect of judgment separately. For the thing that determines whether or not wetness is treated as an internal

[76] Again, insofar as we focus merely on the "relational" aspect of a judgment and the associated concepts of inner and outer, the "quality" of the judgment is undetermined, which is why I include the possible denials of these judgment forms.

or external property is the *context* of the judgment and how *other comparisons* are operating.

For example, suppose I am comparing *this* pavement with wetness. If that is the case, I have compared a particular object with a concept. I can see then, that wetness is (at the moment I make the judgment) an internal mode of this pavement. But if I aim to make a judgment about pavement in general, I implicitly make a judgment about all the possible objects that fall under the concept of pavement. In that case, wetness and the concept of pavement in general are only possibly related if other conditions hold, i.e., like rain.

b. Comparing Representations According to their Quantity: The Concepts of Identity and Difference

A comparison via the concepts of identity and difference determines the "quantitative" aspect of a judgment. Like I said above, the quantitative aspect of a judgment concerns how many things something is predicated of, namely, whether it is predicated of "all" of something, "some" of something, or of a singular thing. Here, I propose that the act of comparison that grounds this logical aspect of a judgment is the act of comparing members of a set in regards to another concept. In other words, a judgment of the form "All *A*s are *B*s," presupposes that I have compared all of the objects that fall under the concept A in regards to whether or not they are similar or different in regards to concept B.

Suppose I wonder how many of the desks in my classroom are left-handed. Kant's theory is that I should compare all the objects that fall under the concept "desks in my classroom" and see if they are similar or different in regards to the concept of left-

handedness. If they all are *identical* insofar as I consider them under this concept, I will judge that *all* the desks in my classroom are left-handed. If, on the other hand, I compare them and notice that some are different than others in regard to the concept of left-handedness, I will judge that *some* of the desks in my classroom are left-handed. Hence, the concept of identity is associated with the universal judgment and the concept of difference is associated with the particular.

Notice, when I compare objects in regards to their identity and difference, this is not to say that I compare the objects that fall under concept A with the set of objects that fall under concept B. Instead, I compare at least two objects *with each other* with regard to whether or not they are similar or different in regards to another concept. I point this out because it explains why there is no concept of comparison associated with the third quantitative aspect of a judgment, i.e., singularity. A singular judgment, i.e., "*this* desk is left-handed" does not involve comparing members of a set of desks to see in what way they are similar or different. Rather, such a judgment simply affirms something of a singular object. Hence, while such a judgment does not presuppose comparison via the concepts of identity and difference, it instead presupposes a comparison via the concepts of agreement and opposition.

c. Comparing Representations According to Their Quality: The Concepts of Agreement and Opposition

The qualitative aspect of a judgment, recall, expresses an affirmation or denial (or "negation" to use Kant's language). A comparison via the concept of *agreement* precedes an affirmative judgment and a comparison via the concept of *opposition* precedes a negative one. This means I compare two concepts to see whether or not they

are both attributable to the same object. If they are, they "agree." If they are not, they "oppose" one another. For example, the judgment *this desk is left-handed* presupposes that I have compared this desk with the concept of left-handedness to judge that it is the case that the concept of left-handedness is indeed attributable to this desk. Likewise, if I judge that it is not the case that this desk is left-handed, I have compared this desk with the concept of left-handedness to judge that it is not the case that the concept of left-handedness is attributable to this desk. This can happen on the same level with a universal judgment—it is just that with those judgments I compare a set of objects with a possible property to see if they are in possible agreement or not. Hence, I can judge that some cats are black, indicating that blackness can be attributed to some cats. Likewise, I can judge that it is not the case that all cats are black, indicating that the concept of blackness is not attributable to *all* cats.

Like I have said, treating the concepts of comparison as separate and hence independent of one another is misleading. The way in which we compare representations will always depend on context. In this vein, different aspects of a judgment might accordingly be highlighted in different judgments. We might judge that *all* men are mortal, indicating concern with the quantity aspect of the judgment or we might judge that men are *mortal*, as opposed to immortal, indicating concern with what kind of predicate is attributable to the subject of judgment. Likewise, we might judge that all men *are* mortal (*period)*, indicating that we want to emphasize that no external conditions are necessary in order for men to be mortal.

There is another difficulty that Kant himself focuses on in the Amphiboly chapter, which stems from the fact that we can use the concepts of comparison in two different ways, leading to ambiguous judgments. It is this possible twofold use of the concepts of comparison that makes transcendental reflection necessary. I now turn to discussing it.

4. The Concepts of Comparison and the "Twofold Relation" of Things to Cognition

Recall that transcendental reflection is the activity of distinguishing whether representations are to be "compared to one another as belonging to the pure understanding or to pure intuition" (A 261/B 317). It is necessary, Kant says, because "things can have a twofold relation to our power of cognition, namely to sensibility and the understanding" (A 262/B 318). Now, Kant does not directly argue for this claim, but it is clear that one of the reasons why transcendental reflection is necessary is that it allows us to distinguish between claims that are genuine *synthetic a priori* truths from those that merely appear to be. As he states, it is a "duty from which no one can escape if he would judge anything about things *a priori*" (A 263/B 319).[77] The point that transcendental reflection serves to identify which *a priori* truths are synthetic is important for my analysis of the paralogisms. The rational psychologists think they infer synthetic *a priori* truths about a thinker that are indeed *a priori* since they are necessarily true, but

[77] It seems then, that transcendental reflection is what Kant later identifies as the "obligation of a philosopher": "it is the utmost importance to isolate cognitions that differ from one another in their species and origin," so we can "securely determine the proper value and influence of the advantage that a special kind of cognition has over the aimless use of the understanding" (A 842/B 870). The "species and origin" of a cognition [*Erkenntnis*] refers to either the understanding or sensibility. Thus, when Kant goes on to say that such a separation and isolation is to obligation of a philosopher, he means that a philosopher is obliged to track the influence of the sensibility or the intellect for the sake of determining how each influences the way we understand the world.

not in fact synthetic, since they do not inform us about the way that one necessarily exists as a thinker.

For now, though, I discuss Kant's point that objects have a two-fold relation to cognition. As Kant says, an act of comparison precedes objective judgments, which means that the comparison is not merely logical but involves comparing the representations of *objects*. As Kant has argued in the first half of the *Critique*, we can represent objects in two ways: as objects of mere thought and as objects of intuition. Of course, Kant argues that knowledge of objects requires both intuitions and concepts (A 51/B 75). But this does not entail that we cannot *consider* an object through concepts alone or through intuition alone. Doing the first would involve abstracting from all of the object's spatiotemporal qualities and focusing solely on its conceptual properties, while doing the second would involve abstracting from all of the object's conceptual properties and focusing solely on its spatiotemporal properties. As we will see, Kant consistently uses the phrase "abstracting from" when he discusses the activity of comparison, implying that we *begin* with an object with both conceptual and spatiotemporal properties.[78] In this sense, when Kant refers to a "pure object of the understanding," we can take him to mean an object insofar as we have abstracted away all of its spatiotemporal properties.

Thus, when Kant states that transcendental reflection is an activity of distinguishing whether representations are to be "compared to one another as belonging

[78] The locution "abstracting from," Kant points out, is importantly different from that of "abstracting something." For example, when we consider the scarlet color of cloth, we *abstract from* the cloth for the sake of focusing on its color" (*The Jäsche Logic*, §6).

to the pure understanding or to pure intuition" (A 261/B 317), he means that it serves to index *how* a pre-judgmental comparison was made, whether by considering an object's conceptual properties or its spatiotemporal properties. Failure to index the cognitive origin of such comparisons is to engage in what Kant calls a "transcendental amphiboly," which is a "confusion of the pure object of the understanding with the appearance" (A 270/B 326). Admittedly, Kant uses the word "appearance" to refer to an object of experience, which has both spatiotemporal and conceptual properties. So we might think this runs against my interpretation, which is that an amphiboly is to confuse the pure object of the understanding with the object considered merely as a spatiotemporal object. But as we shall see, the proper issue at stake is whether or not one makes the proper distinction between the sensible and conceptual aspects of an object. The question is whether or not one admits that in order to think objectively, both senses and concepts are necessary and that neither can be reduced to the other. To fail to transcendentally reflect is to fail to recognize that some judgments arise only after one *abstracts from* an object's spatiotemporal or conceptual properties.

Accordingly, there are two ways to commit a transcendental amphiboly: 1) one can privilege the faculty of understanding, judge what is true of an object's conceptual properties, and take those judgments to be true of an object in general, including as it is exists spatiotemporally, or 2) one can privilege the faculty of sensibility, judge what is true of an object's spatiotemporal properties, and take those judgments to be true of an object in general, including how we conceive of it. Kant accuses Leibniz of the first mistake and Locke of the second. Leibniz "believed himself able to cognize the inner

constitution of things by comparing all objects only with the understanding and the abstract formal concepts of its thinking" (A 270/B 326), and Locke "sensitivized the concepts of the understanding … [by] interpreting them as nothing but empirical or abstracted concepts of reflection" (A 271/B 327). Both err because they do not distinguish the understanding and sensibility as necessary but separate faculties of cognition: "Instead of seeking two entirely different sources of representation in the understanding and the sensibility, which could judge about things with objective validity **only in conjunction**, each of these great men hold on only to one of them, which in his opinion is immediately related to thing in themselves, while the other does nothing but confuse or order the representations of the first" (ibid., Kant's bold).

Recall that when we compare representations, we do so via a pair of concepts. To yield the quantity aspect of a judgment, for example, I compare all the objects that fall under a certain concept (all the desks in my classroom) to see if they are identical or different in regards to another concept (left-handedness). Notice that in this example, it is clear what constitutes the *identity condition*: the desks are identical in this sense *if* they all fall under another concept (that of left-handedness). This condition is what guides my recognition that the desks are *all* left-handed. But this is not the only identity condition that could guide our comparisons: one could be guided by the condition that something is identical to something else if and only if it shares the same space at the same time, which is of course, the condition for numerical identity. Thus, we have two possible conditions that might operate when we compare objects. It is no coincidence that these conditions mirror the twofold relation that things share with cognition: the first expresses the

condition for identity when we compare objects in their conceptual aspects, the second for when we compare object's spatiotemporal ones. *This* is what Kant is worried about in the Amphiboly chapter: the fact that the conditions of comparison change depending on whether one considers an object merely conceptually or merely spatiotemporally.

Of course, there is no risk of confusion with the above example. In everyday comparisons, the conditions are obvious. It is with metaphysical principles, we will see, that the concepts of comparison threaten to misguide us. First though, it will help to recognize the possible different conditions for each pair of concepts. Here, I will present Kant's own distinctions and offer a formal interpretation of them:

a. Inner and Outer

In regards to inner and outer, Kant says:

> In an object of the pure understanding only that is internal that has no relation (as far as the existence is concerned) to anything that is different from it. The inner determinations of a *substantia phaenomenon* in space, on the contrary, are nothing but relations, and it is itself entirely a sum total of mere relations. (A 265/B 321)

This indicates that he is thinking of the conditions for the concepts of inner and outer in the following way:

1. When we abstract from the spatiotemporal properties of an object and consider it merely via its conceptual properties, a predicate is attributable to an object with no conditions added (i.e., as "inner") if that predicate is contained within the concept of the object. Otherwise, the predicate is only attributable to an object hypothetically, i.e., under certain conditions

(and hence is "outer"). Call this the *conceptual condition for inner and outer*.

2. When we abstract from the conceptual properties of an object and consider it merely via its spatiotemporal properties, a predicate is attributable to an object with no conditions added (i.e., as "inner") if that predicate is spatially internal to the other. A predicate is only hypothetically attributable to an object, i.e., under certain conditions, if that property is spatially external to the other. Call this *the spatiotemporal conditions for inner and outer*.

Kant turns to a discussion of substance for an example of comparisons via the concepts of inner and outer. When we consider attributing predicates to a subject, conceptually speaking, we must think of that subject as having something inner, i.e., something that does not depend on its existence on anything external to it, since anything external constitutes a possible predicate of that subject. "Substances in general," as Kant says, "must have something *inner*, which is therefore free of all outer relations… the simple is therefore the foundation of the inner in things in themselves" (A 274/B 330). This is just to say that when we attribute a property to an object, as we do in the judgment "the desk is brown," we necessarily conceive of the object as a substance, i.e., as having an inner foundation to which we can attribute properties.

But notice that to judge something as a substance via the conceptual conditions of inner and outer does not necessarily entail that it is a spatiotemporal substance, insofar as we think of "substance" as something absolutely inner. This is because the

spatiotemporal conditions for inner and outer involve not how we must conceive of an object but rather how it exists. Strictly speaking, when comparing representations spatially, because all spatial objects by definition occupy space, everything is "absolutely" outer and only "relatively" inner, which is to say there is no true "inner" spatially speaking. To say that the determinations of a "*substantia phaenomenon*" (a spatial object) are nothing but relations, as Kant says in the above quote, is to say that the identity conditions of a sensible object must always refer to relational conditions and not merely conceptual ones. In other words, a purely conceptual description of a spatiotemporal object will never serve to uniquely identify a spatiotemporal object. To identify the exact "green vinyl chair" I aim to refer, I must also include information about the relational qualities of that green chair, e.g., that it is the one to the right of the table in the corner in the conference room, and so on. It is only in reference to objects *external* to the object that I can identify it. While we must conceive of an object as having non-relational identity conditions, this does not entail that it will have non-relational identity conditions insofar as we consider it as a sensible object.

Leibniz engages in an amphiboly because he holds that the conceptual conditions for inner and outer are applicable to *all* objects, including those of experience. Hence, he claims to have discovered a *synthetic a priori* principle, namely, that monads, which are absolutely inner in the conceptual sense "constitute the fundamental matter of the entire universe" (A 330/B 274). And Kant makes clear that Leibniz would be right,

> were it not that something more than the concept of a thing in general belongs to the conditions under which alone objects of outer intuition can be given to us, and **from which the pure concept abstracts**. For these show that a persistent appearance in space (impenetrable extension) contains mere relations and

> nothing absolutely internal, and **nevertheless can be** the primary substratum of all outer perception. A 283-84/B 340, my emphases.

In other words, if there is no distinction between a pure object of the understanding and the appearance, then Leibniz would be right that monads constitute the fundamental matter of the entire universe. But that there must be something "absolutely inner" grounding all matter is true only insofar as we "abstract from" the conditions under which objects of outer intuition can be given to us—i.e., from spatiotemporal conditions. When we consider objects in space, we can see that the conditions for inner and outer change and hence that we do not need an *absolute* inner—spatially speaking—for something to count as a substance. Thus, Leibniz misidentified an analytic *a priori* truth for a synthetic one.

b. Agreement and Opposition

In regards to agreement and opposition, Kant says the following:

> If reality is represented only through the pure understanding… then no opposition between realities can be thought, i.e., a relation such that when they are bound together in one subject they cancel out their consequences …. Realities in appearance… on the contrary, can certainly be in opposition with each other and, united in the same subject" (A 264-65/B 320-21).

This passage indicates that Kant is thinking of the conditions for agreement and opposition in the following way:

1. When we abstract from the spatiotemporal properties of an object and consider it merely via its conceptual properties, a property agrees with another if one can attribute both to the same object and such attribution does not result in a logical contradiction. Otherwise, they "oppose" each other. Call this *logical agreement or opposition*.

2. When we abstract from the conceptual properties of an object and consider it merely via its spatiotemporal properties, a property "agrees" with another if one can attribute both properties to the same spatiotemporal object. Otherwise, these properties "oppose" each other. Call this *spatiotemporal agreement or opposition*.

For example, conceptually speaking, the concepts of "black" and "not black" are not both attributable to one object. To say that an object is both black and not black is to express a contradiction. But sensibly speaking, the concepts of "black" and "not black" *are* both attributable to one object, precisely because a sensible object exists in space and time and thus can be black and not black at different times or can be black in one spatial region and not black in another. When we consider that object merely conceptually, however, we abstract from all of its spatiotemporal properties and can no longer attribute opposing properties to it. This indicates that conceptual opposition signifies a contradiction, while sensible opposition signifies a spatiotemporal change in or variation of the same object.

c. Identity and Difference

In regards to the concepts of identity and difference, Kant says:

> If an object is presented to us several times, but always with the same inner determinations (*qualitas et quantitas*), then it is always exactly the same if it counts as an object of pure understanding, not many but only one thing (*numerica identitas*); but if it is appearance, then the issue is not the comparison of concepts, but rather however identical everything may be in regard to that, the difference of the place of these appearances at the same time is still an adequate ground for the numerical difference of the object (of the senses) itself. (A 263/B 319)

This passage indicates the following identity conditions:

1. When we abstract from the spatiotemporal properties of objects and consider them merely via their conceptual properties, two or more objects are identical in regards to a concept if they both have that same property (i.e., if they both fall under that concept). Otherwise, they are different in this respect. Call this *qualitative identity or difference*.
2. When we abstract from the conceptual properties of objects and consider them merely via their spatiotemporal properties, an object is "identical" to another if they are in the same space at the same time. If not, those objects are different. Call this *numerical identity or difference.*

For example, I could compare two water drops for the purpose of forming a judgment about them (A 263-4/B 319-20). The water drops could be identical or different in one of two ways: two numerically different water drops can be qualitatively identical, meaning that they share all the same qualities. Similarly, they could be numerically identical (i.e., the same water drop) if they are always in the same space at the same time. The crucial point to notice is that numerical identity or difference does not entail qualitative identity or difference and vice-versa. Two numerically different water drops can still be qualitatively identical—i.e., they can share all the same properties but exist in different spatiotemporal locations. Likewise, *one* (numerically identical) water drop, can be qualitatively different over time, i.e., it can change its properties and still be the same water drop. Hence, the judgment "water drop B and water drop A are

identical," is amphibolous: it can mean either that the water drops are numerically identical or that they are qualitatively identical.

Transcendental reflection is the activity of recognizing which conditions (those I have classified under (1) or (2)) have guided the pre-judgmental comparisons. Hence, it is to qualify judgments to rid them of any ambiguity, e.g., "water drop B and water drop A are numerically identical," or "the table is a spatiotemporal substance." Of course, under normal conditions we have no problems recognizing the proper meanings of our judgments. It is when the conditions of comparison are used to generate metaphysical truths that they become troublesome. A metaphysician who privileges the understanding, for example, will take it that the condition for qualitative identity constitutes identity conditions *in general* and take it to be true not just about objects of the pure understanding but also about any kind of object, including sensible ones. Hence, we end up with principles like Leibniz's identity of indiscernibles that states that no two objects have exactly the same properties, a principle that would have major metaphysical implications, if true. As Kant notes, Leibniz "believed himself to have made no little advance in the cognition of nature" (A 272/B 328). But as Kant goes on to say, it is clear that a difference in the spatiotemporal location of an object will make a difference in whether or not we judge it to be identical to another object. "Thus," *and this is the crucial point,* "that putative law is no law of nature. It is simply an analytical rule or comparison of things through mere concepts" (ibid.).[79]

[79] So far, I have not yet discussed any of Locke's mistakes based on his amphiboly. We will see that Kant is mainly concerned with Locke's mistaken application of the concepts of identity and difference, which indeed, is the one that manifests itself in the third paralogism. I discuss it below.

Thus, we have full confirmation of my claim above that transcendental reflection will serve to identify when an *a priori* claim is synthetically true or when it is only analytically true. We also see here that the lack of transcendental reflection—that is, the inability to recognize that the understanding and intuition are both necessary for knowledge—prohibited Leibniz from understanding the true significance of those principles he took to ground metaphysics. It is not that principles such as the identity of indiscernibles are not true—they are. They are just not informative of our experience in the way Leibniz thought—they are only analytically true. In the next section I show how the paralogisms can be similarly characterized.

5. Kant's Theory in the Amphiboly as the Key to Understanding the Paralogisms

We are now in a position to see how Kant's theory in the Amphiboly is instructive in understanding the arguments of rational psychology. In section 2, I showed that any adequate account of the paralogisms will not only reconcile the various interpretative issues surrounding them (i.e., whether they are valid, what type of fallacy they commit, and what Kant agrees with as far as their content is concerned), but also explain their connection to transcendental illusion.

I will begin by addressing the last point. I argued in chapter 2 that we must presuppose that the soul, the universe as a whole, and God, exist. We saw there that although the illusion is necessary, we can avoid being deceived by it. The fallacious arguments of rational psychology arise as a result of being deceived by the illusion. My claim here is that the deception is a result of not being in the position to transcendentally reflect. Thus, while transcendental illusion is inevitable for both the transcendental

realist and the transcendental idealist, the latter can stop themselves from making faulty inferences based on it, since they are in the position to transcendentally reflect.

The rational psychologists begin with the transcendental illusion that the soul necessarily looks like an object to us, which is to say, they begin with the assumption that one has a soul about which they can gain further knowledge. This motivates them to consider what types of judgments they could make about such a soul. We will see that they assert judgments about the soul that are expressed in the minor premises of the paralogistic arguments. Because transcendental realists are not in the position to transcendentally reflect, they are not able to recognize that such judgments are ambiguous. Hence, they misunderstand their significance and infer principles that do not in fact inform us about the way one exists as a thinker.

My argument is that each fallacious argument is grounded on a comparison that takes place via a pair of the concepts of comparison. The comparisons can be summarized in the following way:

1. I can compare myself as a thinker with my thoughts. Here, the concepts of inner and outer will lead to the judgment that I am the absolute subject of my thoughts, which is the minor premise of the first paralogism.
2. I can compare myself as a thinker with the activity of thinking. The concepts of agreement and opposition will lead to the judgment that I am a thing that cannot be "regarded as the concurrence of many acting things," (i.e., that I am not many things), which is the minor premise of the second paralogism.

3. I can compare myself as a thinker at one time with myself as a thinker at another time. Here, the concepts of identity and difference will lead to the judgment that I am conscious of the numerical identity of myself in different times, which is the minor premise of the third paralogism.

Each of these minor premises is ambiguous, as I will show. Additionally, it is because the rational psychologists engage in an amphiboly that they are not able to recognize that ambiguity. Hence, they take themselves to arrive at *synthetic a priori* truths about the nature of the soul that are in fact only analytic truths about the transcendental unity of apperception. I will now treat each of the paralogisms in turn. Along the way, we will see how the interpretive issues I detailed in section 2 are solved, namely: that the arguments are properly understood as invalid arguments that commit the fallacy of the ambiguous middle. We will also be in a position to identify which claims of the rational psychologist Kant finds unproblematic.

a. Comparing a Thinker to its Thoughts. The Concepts of Inner and Outer and the First Paralogism: "I am a Substance."

Recall the first paralogism:

> [1] That the representation of which is the **absolute subject** of our judgments, and hence cannot be used as the determination of another thing, is **substance**.
> [2] I, as a thinking being, am the **absolute subject** of all my possible judgments, and this representation of Myself cannot be used as the predicate of another thing.
> Thus I, as a thinking being (soul), am **substance.** (A 348, Kant's emphases)

We can see now how an act of reflection via the concepts of inner and outer grounds the second premise. Kant complicates this premise by providing a complex middle term, "the absolute subject of all my possible judgments and [hence something] that cannot be used

as the determination [or predicate] of another thing."[80] The basic judgment, however, is that I am the subject of thought and not the predicate (of thought or anything else). Hence, the logical form of this judgment is that of a singular, affirmative, categorical one. Here, we focus on the categorical aspect, which is to focus on the fact that subject-hood can be predicated of me, considered as a thinking being, with no conditions added, which is why the premise refers to *absolute* subject-hood. It is not as though I am the subject of my thoughts (and not the predicate) only under certain conditions—I am the subject of my thoughts *period*.

Notice that the second premise then, has presupposed that I have compared myself as a thinker with my thoughts in a way that relied on the *conceptual* senses of inner and outer, which is to say that when I reflect on myself as a thinker, I see that the concept of subject-hood is necessarily contained in the concept of a thinker. So far then, the second premise does not state anything that Kant himself does not agree with. He makes clear in the Transcendental Deduction that a thinker must conceive of itself as the subject of thought: "all manifold of intuition has a necessary relation to the **I think** in the same subject in which this manifold is to be encountered" (B 132). But notice that *Kant* is in the position to transcendentally reflect on this judgment and recognize that it is a truth about me only insofar as I consider myself conceptually.

The rational psychologists, on the other hand, are in no such position. Because they confuse the object of pure thought with the appearance, which is to say that they commit a transcendental amphiboly, they take conceptual truths to be unqualifiedly true

[80] It is unclear why Kant changes the language mid-argument. I proceed here under the assumption that he means to identify the same term in both the major and minor premises.

of *all* objects, including spatiotemporal ones. So when they judge that I must conceive of myself as the absolute subject of judgments, they take it that *I am* the absolute subject of judgments. Given the basic definition of substance that is expressed in the major premise—that anything that is an absolute subject and cannot be thought of as a predicate is a substance—it follows from this reasoning that I am a *substance*.

Transcendental reflection on the second premise reveals the fallacy. If we qualify the judgment "I as a thinker am the subject of all my possible judgments" in a way that identifies the cognitive origin of the judgment, we see that it is true only under certain conditions, namely, the very conditions for discursive thought. Doing so yields a very different judgment: "If I consider myself merely as a thinker, abstracted away from all my spatiotemporal qualities, then I am the subject of all my judgments." Notice that the "absolute" is no longer applicable, since the judgment now expresses a *conditional* truth. *This* premise is not subsumable under the general definition of substance given in the major premise, for the definition refers not to those things that we can think of as subjects under certain conditions, it refers to *absolute* subjects.

We can see now how this interpretation allows us to reconcile Kant's comments about the fallacy supposedly committed by the rational psychologists. Recall that he claims it is a formal fallacy of the ambiguous middle (see A 341/B 399 and A 402-3/B 411). While the argument's form on its surface is valid, it is only so because the rational psychologists have failed to make the significance of the second premise explicit. If they were to recognize that it states merely a conditional claim, they would see that they are not entitled to predicate *absolute* subject-hood to a thinker considered merely as such.

Thus, the ambiguity does indeed lie in the middle term, which refers to that which is "the **absolute subject** of all my possible judgments and … cannot be used as the predicate of another thing." Hence, we can see why Kant classifies it as a fallacy of the ambiguous middle.

Furthermore, we can see why Kant sometimes implies that it is the *category*—in this case, substance—that is being used ambiguously insofar as it is being used "transcendentally" in the major premise and "empirically" in the minor (A 402). The major premise states a definition of substance, which makes no distinction in the way we consider an object. It is, as Kant says, "a pure intellectual concept, which in the absence of conditions of sensible intuition is merely of transcendental use, i.e., of no use at all" (A 403). The major premise, in other words, defines substance "in the absence of conditions of sensible intuition." But the minor premise, *properly expressed*, refers to the discursive conditions under which I can properly think of myself as the subject of my thoughts. Hence, all I could properly deduce about my nature as a thinker would have to be under those same conditions of sensible intuition, which is to say that I could only apply the empirical use of the concept of substance. But that is not what the rational psychologists claim to do; hence, they implicitly rely on two different uses of the concept of substance. It is no wonder then, that Kant's identification of the mistake here—as one that confuses the transcendental and empirical use of a concept—repeats his classification of the amphiboly as a "confusion of the empirical use of the understanding with the transcendental" (A 260/B 316).

My interpretation also reveals that Kant's classifications of the fallacy in the second edition, where he offers a seemingly different account, are consistent with those of the first. In the second, Kant says that,

> 'thinking' is taken in an entirely different signification in the two premises: in the major premise, as it applies to an object in general (hence as it may be given in intuition); but in the minor premise only as it subsists in relation to self-consciousness, where, therefore, no object is thought, but only the relation to oneself as subject (as the form of thinking) is represented. In the first premise, things are talked about that cannot be thought of other than as subjects; the second premise, however, talks not about **things**, but about **thinking**.... (B 411fn, Kant's bold)

Again, transcendental reflection illuminates Kant's remarks here. If we reflect on the second premise (as it is originally expressed), we see that it elides the two possible conditions under which we can compare our representations in regards to the concepts of inner and outer. The second premise does not qualify that it is only true insofar as we consider ourselves as thinkers merely conceptually. Hence, it does not consider a *thing* that thinks, it merely considers oneself as one is engaged in the activity of thinking. Of course, if we so qualify the second premise, we see that it is no longer subsumable under the major premise, since the major premise, as Kant says in this passage, applies to *objects*.

The point that the rational psychologists consider a thinker as it engages in thinking instead of as an object illuminates why Kant says that they consider the proposition I think "taken problematically" (A 348/B 406). This means that they consider the thinker merely hypothetically, i.e., *if* something thinks, it must be the absolute subject of judgments. This is as opposed to considering a thinker assertorically, which would be to consider a thinker as an actually existing thing. If the rational

psychologists admitted that their conclusions describe merely how we must conceive of something in order for it to be a thinker at all and not how a thinker exists, they would recognize their fallacious reasoning.

Transcendental reflection is indeed, the way rational psychologists would recognize their fallacious reasoning, since as I said above, it reveals which *a priori* truths are truly synthetic. Kant repeatedly says that the rational psychologists make merely an analytic claim that they try to pass off as a synthetic one (A 350-51; A 355; A 366). Here, we can see that they do this precisely because they are not in the position to transcendentally reflect.

b. Reflecting on the Activity of Thinking. The Concepts of Agreement and Opposition and the Second Paralogism: "I am Simple."

The second paralogism states:

> That thing whose action can never be regarded as the concurrence of many acting things, is simple.
> Now the soul, or the thinking I, is such a thing.
> Thus etc. (A 351)

Just as with the first paralogism, I propose here that the second premise is grounded by an act of reflection via a pair of the concepts of comparison, in this case, agreement and opposition. The basic judgment of the second premise is that "I cannot be regarded as the concurrence of many acting things," and hence has the logical form of a categorical, singular, negative judgment. Here, we focus on the negation: that I am not something.

The second premise then, has presupposed that I have compared myself as a thinker with the activity of thinking in a way that relies on the *logical* senses of agreement and opposition. When I reflect on myself as a thinker, I see that the activity of

thinking is logically opposed to me being many different things. As Kant says, we must consider the type of activity a composite substance can and cannot engage in: "Every composite substance is an aggregate of many, and the action of a composite, or of that which inheres in it as a such a composite, is an aggregate of many actions or accidents, which is distributed among the multitude of substances" (A 352). In other words, a composite substance—that is, a whole made up of more than one single substance—can, for sure, act. But the actions will be "distributed among the multitude of substances." Suppose I kick a soccer ball. This action is one of an aggregate substance, my body. But the action itself can be considered as a whole of smaller actions each distributed across the different parts of my body. My knee joint moves in a certain way, while my shin another, etc. All of those actions together, as an aggregate, constitute the action as a whole of me kicking the ball. As Kant goes on to say, however, this is not a possible way of viewing the activity of thinking:

> Yet with thoughts, as accidents belonging inwardly to a thinking being, it is otherwise. For suppose that the composite were thinking; then every part of it would be a part of the thought, but the parts would first contain the whole thought only when taken together. Now this would be contradictory. For because the representations that are divided among different beings (e.g., the individual words of a verse) never constitute a whole thought (a verse), the thought can never inhere in a composite as such. (A 352)

William James famously expands on this passage by asking us to suppose that twelve men each have in their heads a different word of a whole sentence. At no point does any individual man have an understanding of the whole sentence. It requires *one* thinker to synthesize the words into the single thought in order to understand the sentence as a whole, i.e., in order to grasp the thought (1950, 160). This reflection is captured in the second premise.

So far then, the second premise does not state anything that Kant himself does not agree with. Again, he makes clear in the Transcendental Deduction that a thinker must conceive of itself as one thing: "it is only because I can combine a manifold of given representations **in one consciousness** that it is possible for me to represent the **identity of the consciousness in these representations** itself" (B 133, Kant's bold). But, just as with the first paralogism, Kant is in the position to transcendentally reflect on this judgment and recognize that it is true about me only insofar as I consider myself conceptually.[81]

The rational psychologists are in no such position. Indeed, they commit the same mistake as I described in the first paralogism, based on the fact that they do not distinguish between the object of pure understanding and the appearance. Two representations can conflict logically and agree sensibly. While it might be true that I must think of myself as simple, this does not indicate how I necessarily exist. I can only conclude that I am simple insofar as I have abstracted from my sensible qualities. As Kant says:

> But the simplicity of the representation of a subject is not therefore a cognition of the simplicity of the subject itself, since its properties are *entirely abstracted from* if it is designated merely through the expression "I," wholly empty of content (which I can apply to every thinking subject). (A 355, my emphasis)

It is possible that even though I must conceive of myself as simple, I could still exist as an aggregate. Kant makes this point in reply to Moses Mendelssohn's "proof of the persistence of the soul" (B 413 ff). In the *Phädon*, Mendelssohn concludes that the soul must be persistent by reasoning that a simple being cannot gradually transform into

[81] As he says, "this principle of the necessary unity of apperception is, to be sure, itself identical, thus an analytical proposition" (B 135).

nothing (2007, 95-96). If the soul is simple, then it must persist and hence be immortal, since a simple entity cannot go out of existence. In reply, Kant launches into a discussion of what modern philosophy of mind refers to as the problems of "fission" and "fusion." The first refers to one mind splitting into two and the second to two minds fusing into one. These problems touch on what is at stake in the second paralogism: they are problems because we think of the mind as a simple entity, i.e., as *one* thing. But, as Kant goes on to argue, even if the soul is one "thing," we can still think of it as a composite in another sense: as made up of several powers or "faculties," i.e., the capacities to think, judge, sense, etc. Thus, we could think of dividing the soul not in virtue of splitting its unified substance but rather its powers or faculties:

> Just as one can think of all the powers and faculties of the soul, even that of consciousness, as disappearing by halves, but in such a way that the substance always remains; so likewise one can without contradiction represent this extinguished half as preserved, yet not in it but outside it …. (B 416)

The idea here seems to be that the mind could be a "simple" substance and still exist as an aggregate—not an "extensive" aggregate—but as an aggregate insofar as it has several powers or faculties. Each of those powers or faculties has what Kant calls an "intensive magnitude," or a degree of reality ("quantum," as he puts it here). Since these powers admit of degrees, they are divisible. But if they can be divided, there is nothing to stop us from thinking of the "extinguished" parts as preserved together. And insofar as those aggregates make up a whole—independent from the original substance—they form a new substance. Hence, there is at least a possibility that even if I must think of myself as a simple substance, my mind is still composite. The point is that it does not follow

from thought alone that just because I must think of myself as simple that I *am*. The argument of the second paralogism is invalid.

c. Comparing a Thinker to Itself at Different Times. The Concepts of Identity and Difference and the Third Paralogism: "I am a Person."

The third paralogism states:

1. What is conscious of the numerical identity of its Self in different times, is to that extent a person.
2. Now the soul is etc.
3. Thus it is a person. (A 361)

Again, Kant does not state the argument completely. Premise two is presumably meant to say that "the soul is conscious of the numerical identity of its Self in different times." At any rate, my argument here follows the same pattern as the others: an act of comparison precedes the second premise, in this case via the concepts of identity and difference. More specifically, one compares the "I" of "I think," to itself, at different times and sees that *all* of those representations are identical.[82] Notice, though, that the second premise refers to *numerical* identity. Thus, the premise implies that I as a thinker am conscious of myself as I exist spatiotemporally, since it is only under spatiotemporal conditions that one can judge numerical identity. Thus, premise 2 does not merely assert that I am conscious of the same "I" representing my thoughts over time, it asserts that I am conscious of an entity that is identifiable in space and time and that I am conscious of *that* entity being identical over time.

[82] Hence, even though the second premise does not explicitly express a universal judgment in its quantity aspect, we can see that the comparison works in the same way it would if preceding a universal judgment: it compares all the representations that fall under "I think" to see if they are similar or different from one another.

In this way, the third paralogism is different from the first two. While the first two focus on the conceptual or logical conditions of comparison, the third focuses on the spatiotemporal conditions of identity (as opposed to the qualitative condition). And this makes sense: for Locke—an empiricist—is the target of Kant's criticism in the third paralogism. Hence, we see here a criticism Kant did not fully develop in the Amphiboly chapter: how Locke engages in a transcendental amphiboly. So far, we have only analyzed Leibniz's amphiboly and have seen that Leibniz's confusion of the object of the pure understanding with the appearance causes him to think that the conceptual, qualitative, and logical conditions of comparison yield information about *any* type of object, including spatiotemporal ones. But as I noted above, this is not the only way to commit an amphiboly—one can do the opposite and think that the spatiotemporal and numerical conditions of comparison yield information about any type of object, including one of the pure understanding. Now we can see how Locke engages in the latter type of amphiboly.

Now, given my analysis of the first two paralogisms, we might expect here an analysis that shows that Locke uses the spatiotemporal conditions of identity and makes a judgment that is not necessarily true in the qualitative sense of identity. The analysis I will present, however, does not follow this pattern. Rather, we will see that transcendental reflection reveals that Locke is not in the position at all to assert numerical identity in the second premise. He only does so because he privileges sensibility (somewhat inconsistently, as Kant notes), and does not notice how that has made him assert a judgment that should be qualified but not in the way he qualifies it. In other

words, the second premise is already a qualified judgment—it notes explicitly that the identity conditions used in the pre-judgmental comparison were those of numerical identity. Those were simply the wrong conditions. From the rational psychologist's standpoint, one is only in the position to compare the identity of the self abstracted from its spatiotemporal properties. Hence, the only judgment the rational psychologist is entitled to is that I am conscious of the same *representation* of myself over time—i.e., that it is always "I" who thinks my thoughts. But that fact in no way entails that I am an entity that is numerically identical over time.

We should take a closer look at how Locke himself reasons. In attempting to solve the problem of personal identity, he distinguishes between a man and a person. A man has the same "successive Body" over time (1979, II, 27, 8). A person, on the other hand,

> is a thinking intelligent Being, that has reason and reflection, and can consider it self as it self, **the same thinking thing in different times and places**; which it does only by that consciousness, which is inseparable from thinking… For since consciousness always accompanies thinking, and 'tis that, that makes every one to be, what he calls *self*; and thereby distinguishes himself from all other thinking things, in this alone consists *personal identity*, i.e., the sameness of a rational Being…." (II, 27, 9)

He goes on to note that it is a separate question as to whether the same person over time is the same *substance* over time. But this question, he says, does not concern personal identity, since personhood is defined as above. His theory, as such, leaves it a possibility that one could change bodies (substances) and still remain the same person over time.

To a certain extent, then, Locke seems to be on solid ground as judged by the standards of transcendental reflection, which would show us that being the same person

over time does not entail being the same substance over time. Locke correctly distinguishes the identity conditions for personhood from those of a substance. What then is Kant's complaint? Kant's complaint is that Locke still does not recognize that even the *numerical* identity of personhood is not possible to judge from the standpoint of considering the thinker merely as such. Although Locke distinguishes between a person and a substance, he still insists that a person is an entity that remains the same over time, *identifiable* as such. That is, a person for Locke is an entity that can be judged to be the same rational *being*, something that is the "same *self* now it was then; and 'tis by the same *self* with this present on that now reflects on it" (III, 27, 9). As such, a person for Locke is not just identifiable from the first-person perspective of self-consciousness, but from a third-person perspective as well, otherwise, one would not be able to *distinguish* a person from all other thinking beings, which Locke insists that we can do.[83]

But from a third-personal perspective, one is *not* in the position to judge whether or not a thinker is the same person over time—that is, whether that thinker is conscious of its numerical identity over time. This is because the thinker itself is not in the position to judge—from self-consciousness alone—whether or not he is the same "being" over time, i.e., the same consciousness. A thinker can merely judge that the "I" of its thoughts is always the same "I." As such, it is not in the position to *infer* that it is a person, i.e., a thing that is conscious of the numerical identity of itself; rather, the identity of the "I" across its thoughts is already built into the fact that it can think thoughts in time. As Kant puts it:

[83] This explains why Kant expresses the third paralogism in the third person perspective rather than in the first-person perspective of the first two paralogisms (A 361).

> The identity of person is therefore inevitably to be encountered in my own consciousness. But if I consider myself from the standpoint of another (as an object of his outer intuition), then it is this external observer who originally considers **me** as **in time**; for in apperception **time** is properly represented only **in me**. Thus from the I that accompanies—and indeed with complete identity—all representations at every time in my consciousness, although he admits this I, he will still not infer the objective persistence of my Self. (A 362-63, Kant's emphases)

So Kant's complaint is that Locke still needs to transcendentally reflect. Namely, he needs to recognize that the numerical condition of identity is not the appropriate one to use when a thinker compares itself to itself over time. Only a comparison via a third-person perspective can judge whether an object is numerically identical over time. Hence, the only judgment that Locke is entitled to is that when I compare myself as a thinker to myself as a thinker at another time, I necessarily take those two thinkers to be the same. But that does not entail that I *am* the same thinker over time. The third paralogism is an invalid argument. Now that we have seen how Kant's theory in the Amphiboly chapter informs the first three paralogisms, we can see how it informs the fourth, even though the fourth has a different form.

6. Reflecting on Our Relationship with Objects. The Concepts of Form and Matter, the Fourth Paralogism, and the Refutation of Idealism: "I am Separate from My Body."

The fourth paralogism, as I have said, has a different form than the first three. Additionally, in the B edition of the *Critique*, Kant adds a chapter that is meant to replace the discussion of the fourth paralogism, the "Refutation of Idealism," which Kant places in a different part of the book (at the end of the "Transcendental Analytic," for reasons that will become clear). This makes it appear as though Kant has disavowed his original discussion. For these reasons, the fourth paralogism is often wholly disregarded in

otherwise good discussions of the Paralogisms.[84] This is a mistake, since Kant's critique of the fourth paralogism is the philosophical heart of Kant's critique of rational psychology. This only becomes clear, however, by first understanding how the ambiguous concepts of reflection motivate the argument.

The argument is as follows:

> 1. That whose existence can be inferred only as a cause of given perceptions has only a doubtful existence:
> 2. Now all outer appearances are of this kind: their existence cannot be immediately perceived, but can be inferred only as the cause of given perceptions:
> 3. Thus the existence of all object of outer sense is doubtful. This uncertainty I call the ideality of outer appearances, and the doctrine of this ideality is called idealism, in comparison with which the assertion of a possible certainty of objects of outer sense is called dualism. (A 366-67).

The first thing to notice is that talk of the thinking I or the soul has dropped out completely. So we must first understand how Kant understands this argument to fit in with the more specific topic of the soul.

His point, as he makes clear in his discussion, is to address the rational psychologist's claim that the soul is separate from the body. The trouble, and the reason why the argument and subsequent discussion gets unwieldy, is that Kant is using that specific issue in service of the broader issue of transcendental idealism—how perceptions in general relate to objects. The broader conclusion—that the existence of outer objects can only be inferred from our perceptions (by judging that the outer object is the cause of those perceptions)—entails a specific conclusion about the soul and body—that the existence of our body can only be inferred from our perception of it. A thinker must *infer* the existence of his or her body; it is not given immediately and necessarily through the

[84] See Allison, 2004; Grier, 2001; Kitcher, 1990; and Strawson, 1975, to name a few.

phrase “I think.” On the other hand, a thinker does not have to infer his or her existence as a thinker through any perception. It is given immediately and necessarily through the phrase “I think.” As a result, the rational psychologist can conclude that because the existence of a thinker does not depend on the existence of a body, the thinker can exist independently of his or her body, and hence, that mind and body are separate.

Kant’s critique is thus aimed at Descartes. It targets the invalid reasoning from the sole phrase “I think,” to the conclusion that the mind and body are separate, which is of course, how Descartes argues in the *Meditations*. First, we get the *cogito*: “I think, therefore I am,” which states that my existence as a thinker is self-validating. Kant does not disagree. Descartes goes on to argue, however, that the *cogito* implies that I can think of the mind as existing separately from the body, and likewise, that I can think of the body as existing separately from the mind. If I am able to think of the mind as existing separately from the body, then, Descartes reasons, the mind and body are “really distinct,” which means that they can exist without each other (1998, 115). And there we have Cartesian dualism.

This would explain why Kant muddies the waters by not focusing merely on the soul in the fourth paralogism. While the Cartesian conclusion is certainly about the soul and how it exists, the reasoning has much broader implications for an entire metaphysical system. To say the soul is independent from the body is not just to make a point about how a thinker exists: it is to make a point about the soul’s relationship with all outer objects. Indeed, it is to say that while my existence as a thinker is *necessary*, the existence of all other objects is only *possible* (since I am only *inferring* that the outer

objects are what cause my perceptions). This explains why the fourth paralogism is connected with both the logical function and the category of modality and furthermore why the replacement chapter—the Refutation of Idealism—falls after the discussion of the principles connected with the category of modality: after the "postulates," which analyze claims of possibility, actuality, and necessity.

The claim that the mind and body are really distinct—Cartesian dualism—is what Kant refers to as skeptical idealism,[85] which denies that we cognize "existence of external objects of sense" through "immediate perception" and "infers from this that we can never be fully certain of their reality from any possible existence" (A 360). (This also helps us understand the name of the replacement chapter: the "Refutation of *Idealism*"). To show that the Cartesian reasoning is invalid is not only to critique Cartesian mind-body dualism. It is to critique the very metaphysical system that Descartes thinks follows. Indeed, the critique reveals what it means to be a transcendental idealist.

It is crucial to recognize the role the concepts of comparison play in the argument about mind-body independence and the broader argument about idealism. The concepts at work here are *matter* and *form*, and they are, Kant tells us, the concepts that "ground all other reflection, so inseparably are they bound up with every use of the understanding" (A 266/B 322). This is unsurprising given what I just said: the use of these concepts in judgment grounds not only judgments about ourselves but the whole

[85] See B 274. The other type of idealism according to Kant is "dogmatic idealism," which he associates with Berkeley. He takes himself to have proven Berkeley wrong with his argument in the Transcendental Aesthetic.

nature of reality. Kant continues to note the several ways in which the concepts of form and matter can be used, depending on the context:

> [Matter] signifies the determinable in general, [form] its determination (both in the transcendental sense, since one abstracts from all differences in what is given and from the way in which that is determined). The logicians formerly called the universal the matter, but the specific difference the form. In every judgment one can call the given concepts logical matter (for judgment), their relation (by means of the copula) the form of the judgment. (A 266-67/B 322/23)

"Matter," then—on the most general level—refers to anything that is "determinable," "form," its determination. This will mean different things depending on whether the context is metaphysical or logical. In a metaphysical context, the distinction harkens back to Aristotle's discussion of form and matter in *Metaphysics*, where he states that "by the matter I mean, for instance, the bronze, by the shape the plan of its form." He goes on to discuss what scholars now call "prime matter" (although Aristotle himself rejects the idea): "by matter I mean that which in itself is neither a particular thing nor of a certain quantity nor assigned to any other of the categories by which being is determined." Matter, in other words, is just whatever is left once we remove all possible predicates: "when all else is taken away evidently nothing but matter remains." It is, in this sense, characterless. "Therefore, the ultimate substratum is of itself neither a particular thing nor of a particular quantity nor otherwise positively characterized; nor yet negatively, for negations also will belong to it only by accident" (*Metaphysics*, Book VII (Z), 1029a.1-3).[86] For Aristotle, this notion of "prime matter" is pure potentiality. It is

[86] Aristotle 1984, 1624-1625.

form that gives something its actuality, which reinforces Kant's connection of form with the modal concept of actuality and matter with the modal concept of possibility.[87]

In a logical context, as Kant points out, the matter—i.e., what is determinable—is the "universal," and the "specific difference" the form. A universal category in logic, e.g., "dog," that can be made more specific (note the term) by noting its species, e.g., "greyhound." The species represents the specific difference between it and others of the same universal category. Just as with the other concepts of comparison, Kant associates matter and form with a logical aspect of judgment, in this case, that of modality. Here, we can see why Kant would associate the concept of "matter" with the modal category of possibility and "form" with the modal aspect of "actuality." The universal concept expresses the possibilities of all the different ways a specific entity can actually exist. Likewise, in a logical context, Kant himself refers to the content of judgments as the "matter" and the relation between judgments their "form" (*Jäsche Logic*, §18).

These differences indicate that, just like the other pairs of the concepts of comparisons, there are different conditions under which we compare representations according to these concepts:

a. When we abstract from spatiotemporal conditions and consider whether one representation is logically prior to another, the first is matter, the second form. Call this the *logical condition for matter and form*.
b. When we abstract from conceptual conditions and consider whether one representation is prior to another, if one is necessary for the sake of the

[87] Thanks to Jill Buroker for helping me get clear on these relations.

> other's existence, the first is the "form" and the second the "matter." Call this the *sensible conditions for form and matter.*

When one abstracts from spatiotemporal conditions, we see that matter must precede form, since from the standpoint of pure thought one must first have something to think *of*, only after which it can be determined by taking on a particular form. The logical "matter" comes first, the logical "form" after. But as we have seen, it does not follow from this that spatiotemporally, matter will precede form. Indeed, in this case, Kant offers a positive argument for the opposite claim that form precedes matter. In the "Transcendental Aesthetic," Kant argues that space and time are necessary conditions for the experience of objects, which means that they are independent of the matter located in them. This means that for sensible objects, form can "precede" matter.

Again, because Leibniz commits a transcendental amphiboly, he is unable to recognize that the first condition does not apply universally, regardless of how we conceive of things:

> On this account [the account that asserts that matter precedes form], Leibniz first assumed things (monads) and an internal power of representation in them, or order subsequently to ground on that their outer relation and the community of their states (namely of the representations) on that. Hence space and time were possible, the former only through the relation of substances, the latter through the connection of their determinations as grounds and consequences. (A 267/B 323)

Leibniz did not recognize that judgments about form and matter are ambiguous. He took it that matter precedes form, since it must on a purely conceptual level. So the matter—i.e., the substance, monads for Leibniz—comes first, space and time—the form—only after and only as a result of the relationships among substances. Thus, we have Leibniz's relational theory of space and time.

The comparison via the concepts of form and matter is what ultimately grounds the second premise of the fourth paralogism, which states that the existence of all "outer appearances cannot be immediately perceived, but can be inferred only as the cause of given perceptions." It means that we can only infer, as opposed to immediately perceive, the existence of outer objects. We infer their existence because we assume that outer objects are what cause our object-like perceptions. I perceive a desk in front of me now and I infer that the desk itself exists as the cause of that perception. But here the old Cartesian point becomes clear: on this reasoning, there is a "veil of perception" between me and the desk. I cannot access the desk except through my perceptions of it, and therefore, I can never perceive the desk immediately. Thus, I cannot be certain of its existence, I can only, perhaps reasonably, infer its existence from the assumption that it causes my desk-like perceptions.

In this case, then, a thinker compares his relationship with outer objects, which here means anything that is not the thinker considered merely as such. One recognizes that one is conscious of oneself and conscious of outer objects in different ways: consciousness of oneself—what Kant calls "apperception"—is immediate, while consciousness of outer objects is not. In the very activity of thinking, one is aware of one's existence—one's existence necessarily follows from the activity of thinking. Consciousness of objects, however, is mediated by perception. This immediacy of self-consciousness entails the judgment that in regards to my relationship with outer objects, consciousness of myself comes first, only after which I can perceive outer objects. This is what Kant means when he says that perceptions are "properly only a *determination* of

apperception" (A 368, my emphasis). One must think of one's own consciousness as determinable, i.e., as the "matter," and one's perceptions as determinations, i.e., as the "form." One's consciousness is determined by our perceptions. But this means that the existence of outer objects is not immediate the way my self-consciousness is, but rather that the relation between me and those outer objects is mediated by my perceptions.

Now we are in a position to see the ambiguity in premise two. That apperception is determinable and perceptions are what determine it, is true only insofar as one compares a thinker with external objects under the logical condition of matter and form. But once we modify the judgment in this way, it tells us nothing about the actual relationship between me and other objects; it does not tell me, for instance, that the existence of outer objects is doubtful, only that I am compelled to think of my existence as necessary and that I am not so compelled to think of outer objects when I consider things from thought alone.

The Refutation makes a related point but in the form of a positive argument. While the fourth paralogism launches a mere skeptical argument against Cartesian dualism, the replacement chapter aims to refute Cartesian idealism and show that one cannot be conscious of one's own existence in time without the perception of outer objects. It aims to prove that we would not be apperceptive without first or simultaneously perceiving outer objects. Inner experience, as Kant puts it, presupposes outer experience.

The argument begins with a Cartesian insight that I am conscious of my existence as determined in time. "All time determination," Kant continues, "presupposes

something persistent." Kant's point here seems to be that I can only be conscious of my existence in time if I have some resting point by which to measure the passage of time. And the only thing that can measure the passage of time is change. But change is only measurable if it is dependent on a substratum for measuring changes. Hence, my consciousness of my own existence in time presupposes something persistent. Furthermore, the persistent thing "cannot be something in me, since my own existence in time can first be determined only through this persistent thing" (B 275-76). In other words, I or something in me, cannot be the persistent thing by which I measure my own existence, since it is my very persistence that is in question. Hence, I can only be conscious of my own existence as determined in time if I perceive an outer thing persisting, i.e., an outer object. Hence, my consciousness of my own existence—the consciousness of my existence that I immediately perceive from the very act of thinking—is only possible if I perceive outer objects first (or simultaneously).

This argument does not prove that my existence depends on outer objects per se; it is not to provide knowledge that my mind would not exist without my body. Rather, it aims to show that I would not have consciousness of my existence without the perception of outer objects. I will not discuss the soundness of this argument or its intricacies here. I wish merely to note that it picks up on the concepts of the determinable (matter) and the determination (form) and shows that one will come to different conclusions depending on how one considers things, whether purely conceptually or spatiotemporally. In the former, we must think of self-consciousness as prior to and determined by our perceptions. This seems to imply that I can infer certain modal judgments merely from

the phrase "I think": that I exist necessarily and that outer objects only exist as a possibility. These judgments are in a way true. They are only so, however, insofar as I index them to pure thought. But to so index them is to admit that they do not express necessary truths from the spatiotemporal standpoint. And of course, Kant's refutation aims to prove that they are not true insofar as we consider ourselves in the spatiotemporal world.

We can see now why the fourth paralogism is the most important. It proves that all attempts at coming up with purely conceptual truths about the self are idle, including those expressed in the first three paralogisms. Transcendental reflection applied to the concepts of form and matter, in other words, would stop the other arguments short. And indeed, as I have noted, this is Kant's point in the "Transcendental Aesthetic." Seen this way, Kant's critique of the rational psychologist is a natural implication of his transcendental idealism.

7. The "Transcendental Topic" of the Soul

Kant states that armed with the lessons of the Amphiboly we can create a "transcendental topic," or a "doctrine that would thoroughly protect against false pretenses of the pure understanding and illusion arising therefore by always distinguishing to which cognitive power the concepts properly belong" (A 268/B 324). We could think of such a doctrine as what Kant calls the "discipline" of pure reason in regards to the soul. Pure reason, Kant says, needs a discipline, in the sense that it needs to be "disciplined"—i.e., kept regulated and in check so it does not stray into error. As Kant says, pure reason "so badly needs a discipline to constrain its propensity to

expansion beyond the narrow boundaries of possible experience and to preserve it from straying and error that the entire philosophy of pure reason is concerned merely with this negative use" (A 711/B 739). Such a doctrine, in other words, is not properly a "positive" one insofar as it documents the ways in which we can positively think of the soul. Rather, it focuses on the ways in which the pure intellect alone should not inflate its conclusions. It is furthermore limited by the idea that even these positive ways of thinking of the soul do not express absolute truths about how the soul *is*, but rather express ways in which we can legitimately think of it so long as we simultaneously recognize the limits of such claims. Truths about thinkers will then have their limitations built into their articulations, guarding against any temptation to mistake their significance.

As we have seen, the paralogisms are examples of the "false pretences of the pure understanding," i.e., examples of what happens when we privilege the intellectual to the exclusion of everything sensible in regards to the soul. Kant's suggestion that we formulate a doctrine to protect ourselves against such false pretences can be applied to our reasoning about the soul by "always distinguishing to which cognitive power the concepts properly belong." Here, I offer what I think would be involved in such a doctrine of the soul.

Using the category of "substance" as an example, we can see that for Kant, there are two legitimate ways in which one can think of oneself as a substance. First, one can think of oneself as a substance in a merely logical sense, insofar as one is necessarily the subject of predication. Second, one can think of oneself as a substance in the sense that

one is a persistent, permanent object over time, and hence an entity that remains the same while one's properties change. This, of course, refers to one insofar as one is embodied. These two ways are related, as the first is what Kant calls the use of substance as a "logical function" and the second is the use of substance as what Kant calls a "schematized" category.

One of Kant's criticisms of rational psychology is that it does not recognize that the category of substance has "no objective significance at all unless an intuition is subsumed under [it]," and that without such an intuition, the category of substance is "merely [a] function of a judgment without content" (A 348-9). This reiterates the point that the concept of substance only arises as meaningful when in the context of attempting to understand different aspects of our experience, the investigation of which is already situated in a spatiotemporal context. To cut concepts off from this context is to cut off any possibility of subsuming an "intuition" under them, hence barring them from having objective significance.

To say that the rational psychologist's definition of substance is a mere "function of a judgment without content" is to say that it captures merely the form of a logical function of judgment. Substance, according to Kant, is just the category that is rooted in the logical function of subject-hood, hence the logical function of the category of substance is just the logical category of subject-hood—i.e., the *A* in the logical form "*A* is *B*." The category of substance is the logical function of a subject when this function

determines something in space and time.[88] When we think of "substance" in a purely conceptual way, without its application to spatiotemporal experience, we think merely of a logical subject. Mere logical subject-hood tells us nothing about what in the world are actual substances; it does not determine what in our experience *is* a substance.

How does the concept of substance, according to Kant, get its spatiotemporal content? Kant offers us a "schema" of when it does so: when something persists or is permanent in space and time, we can consider it to be a substance (A 182/B 224). In this sense, we can think of ourselves and others as "substances," i.e., as having persistent identities over time. Indeed, this is precisely the way we think of ourselves and others. We should notice, however, that this is a far cry from the rational psychologist's conclusion that we are a substance in a way that implies immateriality. First of all, thinking of ourselves as substances in this sense is essentially tied to the fact that our bodies are physical objects that remain relatively persistent and permanent over time. This is not to say that our notion of ourselves as substances in this sense refers to only our bodies; it is to point out, rather, that without our bodies we have no sense of what we are referring to. Our designation of ourselves as substances, in this sense, properly acknowledges the contributions of both our intellect and sensibility. This informs us of Kant's *metaphysics* of the self: namely, that, theoretically speaking, there is no

[88] The categories, as Kant says, "are concepts of an object in general, by means of which its intuition is regarded as **determined** with regard to one of the **logical functions** for judgments" (B 128-9). He also says later that "the same function that gives unity to the different representations **in a judgment** also give unity to the mere synthesis of different representations **in an intuition**, which, expressed generally, is called the pure concept of the understanding" (A 79/B 104-5).

"metaphysics" of the self insofar as we consider ourselves absolutely apart from our material existence.[89]

These different ways of thinking of the soul as a substance are manifest in what Kant calls the "schema" of the idea of the soul. As I mentioned in the last chapter, a "schema," in this context, is a rule that regulates how we should proceed given different purposes that we have. I also pointed out that in the Transcendental Analytic, Kant says a "schema" is a rule that allows us to apply a pure concept of the understanding to existing spatiotemporal objects (A 138/B 177). Since the soul is not such an object, Kant obviously means to use the word "schema" in a different sense here. And as he says, in this case, we have *only* a schema "for which no object is given… but which serves only to represent other objects to us, in accordance with their systematic unity, by means of the relation to this idea" (ibid.). So the schema of the soul is a rule that allows us to represent other objects in a systematically unified way.

The schema of the soul, Kant says, is to "connect all appearances, actions, and receptivity of our mind to the guiding thread of inner experience **as if** the mind were a simple substance that (at least in this life) persists in existence with personal identity, while its states—to which the states of the body belong only as external conditions—are continuously changing" (A 672/B 700). In other words, we can go ahead and treat the

[89] This is why, in my view, commentators who attempt to posit a Kantian metaphysics of the self with the assumption that it can be grounded on the conceptual requirements of the "I," are misguided. This applies even to Colin Marshall, who distinguishes between "metaphysics" and "Metaphysics" of the self. He describes the latter as the dogmatic metaphysical positions Kant rejects but argues that Kant can accept something of the former, which is supposedly a more innocuous metaphysical stance (Marshall, 2010, 3-4). My interpretation suggests that a theoretical view (as opposed to a practical view) of the self that isn't explicitly and essentially tied to our sensible and hence spatiotemporal experience of ourselves as material objects has no traction for Kant.

soul as if it were a substance, with persistence and identity across time. Given that this is normally how we treat ourselves and others, Kant is confirming that we do not have to abandon our common-sense practices and attitudes. The "as if" reminds us that we should not interpret our practices and attitudes to have a metaphysical significance that they do not have (see also A 771-72/B 799-800).

One implication of Kant's point that this is *only* a schema—i.e., one without an object—is that it does not theoretically *determine* how we ought to think of the soul. This is an important difference between the schema of the soul and a schema in Kant's original sense. The schema of substance, for example, is a rule that helps us determine which spatiotemporal objects are substances. The schema of the soul, on the other hand, does no such thing: it merely informs us how to proceed *given certain purposes we have*. We can treat the soul as a simple substance unified across time, then, not because it is, but because treating it as such helps us unify and extend our knowledge about the world. The significance of a schema in this sense is that it gets its grip not from theory but from *practice*. And indeed, it is our moral practices that really motivate the schema of the soul.

In order to take responsibility for our own actions, and indeed, in order for us to think of ourselves and others as properly held to moral standards, we must think of ourselves and others as beings with identities across time, even while our bodies continuously change. We will, however, be continually tempted to assert that these truths must be unqualifiedly true, and claim that because we must conceive of the soul as substance, it *is*, or that because we must conceive of objects spatiotemporally, everything

fundamentally real must be traced back to space and time. What Kant means when he states that we cannot know the "thing in itself" is that we should not take our claims to be unconditionally true.

Thus, so long as we remember that the schema of the soul is only significant in the context of our practices, it will not conflict with any theoretical knowledge. Indeed, the schema will support a higher goal—the unification of reason—or the idea that all the different ways in which we reason about ourselves and the world ultimately form a unified, holistic system. Kant is quite clear on this point in the "Appendix to the *Transcendental* Dialectic," where he says, "If merely regulative principles are considered constitutive, then as objective principles they can be in conflict; but if one considers them merely as **maxims**, then it is not a true conflict, but it is merely a different interest of reason that causes a divorce between ways of thinking. Reason has in fact only a single unified interest, and the conflict between its maxims is only a variation and a reciprocal limitation of the methods satisfying this interest" (A 666/B 694). In other words, if we take the schema of the soul to determine how we exist as thinkers, then it will in fact conflict with other commitments we have, namely, our scientific investigation of the world. But if we recognize that our true interest is to unify both our theoretical and practical commitments, we see that the schema of the soul is partly what allows us to do that.

THE ILLUSION OF A MATERIAL SELF: THE ANTINOMY OF THE SOUL

CHAPTER 4

Shortly after his critique of rational psychology, Kant remarks that reasoning about the soul effects a "one-sided illusion." "It is remarkable," he says, "that the transcendental paralogism effects a merely one-sided illusion regarding the idea of the subject of our thought, and for the opposite assertion there is not the least plausibility forthcoming from the concepts of reason" (A 406/B 433). This is to say that when reasoning about the soul merely from the phrase "I think," we are tempted only by the rationalist conclusion that the soul is a simple substance, identical over time, independent from other objects, and hence immaterial, incorruptible, and immortal (A345/B403). To call the soul a "one-sided" illusion instead of a "two-sided" illusion is to claim that an empirical view of the soul—perhaps the view that the soul is material, perishable, and mortal—does not tempt us when we consider what the soul must be like merely from the phrase "I think."

In this chapter, I argue that from another perspective, from which we view the soul as an external object, the soul is a *two*-sided illusion. Namely, when we consider what the soul must be like from the point of view of what kind of objects fundamentally constitute the universe, instead of from the point of view of self-consciousness (the perspective of the paralogisms), we are inclined to think the soul must be *both* an immaterial substance, simple, and imperishable, *and* a material substance, perishable, and mortal. An antinomy or contradiction lurks in our reasoning about the soul, and hence,

we can apply to it Kant's arguments about how to resolve such contradictions—topics he addresses in the chapter called the "Antinomy." Hence, Kant's dissatisfaction with the rationalist's and empiricist's attempts to answer cosmological questions also applies to psychological questions, and thus, his suggested resolution to the former also applies to the latter.

Applying the Antinomy chapter to the soul is not new. All commentators recognize that Kant's discussion of freedom in the Third Antinomy is crucial for his discussion of the ethical dimensions of the self. I discuss it in the next chapter. Here, I argue for the under-appreciated point that the "second antinomy," although explicitly about the composition of the universe, is also relevant to Kant's discussion of the self.

Noticing this is important for two reasons. First, it reveals that despite what some commentators claim, Kant's arguments about the soul or the self have more than historical importance.[90] Most contemporary philosophers are no longer invested in proving the soul to be immortal or imperishable; indeed, a lot of contemporary philosophy of mind has drifted toward the opposite view that the soul, or self or mind, is reducible to material substance or can be eliminated in favor of talking about the states of the brain. Recognizing that for Kant the soul is a two-sided illusion grants him his

[90]According to W. H. Walsh, for instance, Kant's critique manages only to strike down a now unappealing and dead view of the soul (1975, 174). Kitcher correctly argues that the Paralogisms have more than historical import and that Kant's critique of rational psychology also applies to claims of major contemporary philosophy (1990, 182). She even emphasizes Kant's point in the second paralogism that both the rational psychologist and the empirical psychologist can make equal and incompatible claims regarding the simplicity of the soul. As she says in reference to Kant's remark at B 415fn, "If Rational Psychologists are permitted to argue for the simplicity and immateriality of the soul by claiming that they do not see how a material substance could realize the unity of thought, then materialists would be free to employ the same strategy to 'establish' the opposite conclusion. Since the latter do not understand how an immaterial substance could realize the unity of thought, they may claim that the soul is material" (ibid., 202). Though Kitcher does not connect this with the second antinomy, doing so captures this exact conflict.

justified place at the contemporary table and shows us that his critique of metaphysics applies to us still. Second, my analysis illuminates Kant's own commitments. His positive theory of the self—specifically his claim that a human being cannot be exhausted by any empirical description—depends on his showing that we cannot legitimately claim that the soul is nothing but a material body.

In section 1 I introduce the Antinomy chapter and explain how its structure differs from that of the Paralogisms. In section 2 I offer an interpretation of the second antinomy and in section 3 I sketch Kant's resolution to it. In section 4 I present textual evidence that Kant connects the soul with the second antinomy. While Kant obviously thought that the resolution to the second antinomy had implications for how we view the soul, there is textual evidence for the stronger claim that thinking about the soul generates its *own* antinomy. In section 5 I offer such an antinomy and in the final section I offer a resolution to it. The resolution reveals what Kant is and is not committed to regarding the self, the topic of the next chapter.

1. The Antithesis of Human Reason

Kant says that the paralogistic arguments "effect a one-sided illusion," meaning that they tempt us to believe only the rationalist's unverifiable view of the soul. The arguments of the Antinomy chapter, however, effect what I call a "two-sided illusion," meaning that they tempt us to believe two contradictory views about the nature of the universe, both of which seem necessitates by reason but only one of which can be right. "Here," Kant says referring to the investigations of the Antinomy, "a new phenomenon of human reason shows itself, namely a wholly natural antithetic … which guards reason

against the slumber of an imagined conviction" (A 407/B 434). When we reason for unconditioned truths about the universe, we are lead to contradictions that, because they seem irresolvable, do not allow us to be convinced of one view or another.

Kant explains the asymmetry between the two investigations—those of rational psychology and cosmology—by noting that the first endeavors to explain something not given in intuition—the soul—while the second endeavors to explain something at least partially given in intuition—the universe (A 408/B 435). Of course, the universe *as a whole* is not presented to us in intuition, but the antinomy arguments are designed to explain different aspects of our experience of the universe by appeal to unconditional truths about how it exists as a whole (A 415-19/B 443-47). The introduction of intuitive content leads to contradictions because the empiricists now have arguments to make. They remained silent on the question of rational psychology only because they think we can know nothing from "pure" reason alone. Concerning the universe, however, both the rationalist and the empiricist can offer arguments. More importantly, they can both offer *reasonable* arguments—hence the contradictions.

The Antinomy chapter, just like the Paralogisms, proceeds in four parts, each addressing a different "given" aspect of our experience of the spatiotemporal world: 1) space and time, 2) matter, 3) the fact that certain events, objects or states of affairs "arise" or come about and 4) the fact that some things are dependent on other things (see A 415-19/B 443-47). These topics comprise the first, second, third, and fourth antinomies respectively.[91] The chapter proceeds by attempting to find the "unconditioned" for each

[91] I refer to Kant's chapter with the capitalized "Antinomy" and the specific arguments as "antinomies."

of these given conditions, i.e., by attempting to identify what ultimately must be true for each of these aspects to be given to us.[92] The rationalist, who privileges the conceptual, will approach such questions by first considering what must be true from concepts alone and then applying this knowledge to the content of our experience. The empiricist, who privileges the sensible, will approach such questions by first considering what is true given our spatiotemporal experience, period. Both parties offer arguments and conclusions about what must be the case to explain each of the four aspects, represented by the "thesis" and "antithesis" respectively.

For example, when we ask about what ultimately conditions our experience of space and time, the topic of the first antinomy, the rationalist argues that time must have a beginning and that space must be bounded (A 426/B 454). The empiricist, on the other hand, argues that the universe must be infinite in regards to space and time, i.e., that the world has no beginning and no bounds in space (A 427/B 455). Both arguments rely on showing that if the opposite were true we would not be able to explain the given aspect of our experience. The chapter gets its name from the fact that the thesis and antithesis contradict each other, i.e., they are "antinomial."

The structural difference between the Paralogisms and the Antinomy explains why Kant says that pure reason is not properly antithetic, but that human reason is (A 743/B 771 and A 407/B 433). When engaged in reasoning "purely" about putative metaphysical objects that are not presented to us in our spatiotemporal experience (e.g.,

[92] The content of the antinomies does not come out of the blue for Kant. It mirrors Christian Wolff's discussion of cosmology, particularly his discussion of "how we obtain our concept of the world" and "how our world is constituted" (Watkins 2009, 37).

the soul), only rationalism is compelling. But we are *human* reasoners, embodied in a spatiotemporal world. Thus, we cannot approach all philosophical questions through thought or reason alone. We are compelled to consider what ultimately must be the case given *our* type of experience. Hence, while pure reason itself generates no contradictions in the search for ultimate conditions, human reason does.

Since both the empiricist and rationalist have reason on their side, there is no clear way to defend one over the other. One can proceed unreasonably and insist that all things are conditioned by space and time or that space and time themselves are ultimately intellectual concepts, but to do so is to be dogmatic. This is why Kant says that the antinomies might push one to become a dogmatist, i.e., a philosopher who unreasonably insists on the truth of certain propositions, or a skeptic, i.e., a philosopher who claims that because reason appears to lead to contradictions, we must remain skeptical of any metaphysical propositions (A 407/B 434).

My argument in this chapter is that though the Antinomy explicitly addresses questions of the universe, it can also be applied to questions about the self insofar as we consider the self as in or as part of the physical universe. The easiest place to see the application is in regards to the second antinomy, which concerns the ultimate conditions of matter. Unfortunately, the second antinomy is notoriously difficult to interpret. I offer my interpretation of it in the next section.

2. The Second Antinomy

As I said above, an antinomy addresses an aspect of our experience and asks what ultimately has to be the case for us to experience it. In the second antinomy, the conditioned aspect at issue is the existence of matter. As Kant says:

> reality in space, i.e., **matter**, is likewise something conditioned, whose inner conditions are its parts, and the parts of those parts are the remote conditions, so that there occurs here a regressive synthesis, whose absolute totality reason demands; and that cannot occur otherwise than through a complete division… (A 413/B 440, Kant's emphasis).

In other words, we ask what the ultimate conditions must be in order for matter to exist as reality in space. One condition is that it has parts. A material object, as a whole, depends on the materiality on the parts of which is it constituted. But to notice this is insufficient, since we are in the game of identifying the *ultimate* conditions of our experience. It is not enough to say that a material object depends on its parts, since those parts must be conditioned by parts too, and so on. Kant thinks that we can arrive at two possible conclusions: first, that "every composite substance in the world consists of simple parts, and nothing exists anywhere except the simple or what is composed of simples" (A 434/B 462) or second, that "no composite thing in the world consists of simple parts, and nowhere in it does there exist anything simple" (A 435/B 463). These are the conclusions of the "thesis" and "antithesis," aligned with the rationalist's and empiricist's positions respectively.

We might think that the second antinomy is a debate about the existence of physical atoms. This interpretation would have us believe that the rationalist conclusion

is that physical atoms must exist.[93] But this is not compatible with Kant's own designation of the arguments. He clearly aligns the antithesis, the claim that there is nothing simple, with the position of an empiricist, specifically Epicurus, who was an atomist (A 471/B 499). Likewise, he aligns the thesis, "nothing exists anywhere except the simple or what is composed of simples," with the rationalist, the most obvious candidate being Leibniz, who did not believe in physical atoms. Any interpretation should be compatible with these facts.

Thinking of the second antinomy as a debate exclusively about matter is misleading for another reason, as noted by Grier. She argues that while the antithesis argument does end up being about matter or extended substance (because of certain assumptions the empiricist makes), the thesis cannot be about matter. First, Kant never uses the words "matter" or "extended substance" in the presentation of the rationalist's argument, instead favoring terms such as "composite substance" or "substantial composite" (2001, 198). More importantly, the rationalist argues by *abstracting from* the conditions of space and time (2001, 199), which is precisely what we should expect given the way Kant later characterizes the position of rationalism (A 853-4/B 881-2). Hence, the rationalist's conclusion that "every composite substance in the world consists of simple parts, and nothing exists anywhere except the simple or what is composed of simples" is not *exclusively* about matter. It is about composite substance in general, of which matter is a possible example. In this sense, we can reconcile Kant's comment that

[93] As Grier notes, this view is to be found most obviously in Al-Azm, in *The Origin of Kant's Arguments in the Antinomies* (1972, 52) (Grier 2001, 201).

the second antinomy addresses the existence of matter by interpreting it to be possibly but not exclusively about matter.

The thesis argument proceeds by indirect proof, asking us to assume that composite substances do not consist of simple parts and showing that this assumption leads to a contradiction. If you remove "in thought" all the parts of any composite substance you will have nothing left over, in which case no substance would be given. Thus, either it is impossible to "remove all composition in thought" or it is the case that a substantial composite consists of simple parts, which is to say that the initial assumption is false. The first option is not viable according to the rationalist because *substance* is just something that exists without any necessary relation to anything else. Composition is thus a contingent or external relation, meaning that if we are dealing with a composite *substance* we should have something—i.e., substance—"left in thought" when we remove all composition.

This argument can apply to matter. Kant offers such an application, though as we will see, it yields only the conditional conclusion that *if* matter is considered a substantial whole, then simples must exist. When we attempt a complete division of matter, Kant says, the "reality of matter disappears either into nothing or else into that which is no longer matter, namely the simple" (A 413-14/B 441-42). In other words, either matter is not substantial at all, which would indicate that our experience of it as substantial is illusory, or it could disappear into that which is "no longer matter," i.e., the simple. The latter implies that *if* matter is a substantial composite, there must exist immaterial simples in the universe, e.g., Leibnizian monads.

We must clarify here what the thesis argument means by "composition." Importantly, to say that "nothing exists anywhere except the simple or what is *composed* of simples" is not to say that simple constituents somehow form an aggregate in the way a pile of bricks form a house.[94] In other words, it is not as if a whole bunch of monads are piled together in the same space to form an aggregate. Rather, a substantial whole depends on simples in the same way space arises as a result of certain relations among non-spatial monads for a Leibnizian. On this view, while a single monad is not spatial or extended, a "whole of monads," in virtue of the relationships among individual monads, is extended. This is the sense in which a "composite" whole is "composed" of simple entities for the rationalist.[95] A whole is "composed" of simples in the sense that the existence of the whole depends upon relationships between simple, albeit immaterial, entities.

Given what I have said so far, we can see that the rationalist's conclusion has both a broad and narrow interpretation. The first concludes that a composite substance—without regard to whether that substance is material or immaterial—ultimately depends

[94] This metaphor causes Bennett to misinterpret the antinomy. As he says, "Kant must mean … that the existence of a substance does not depend upon any putting-together of parts to compose that substance" (1974, 164). Bennett seems to suggest here that the only way to understand what Kant means by calling a substance "simple" is that it is non-composite, where "composite" means it consists of parts that are "put together." As I argue, "composed" means something quite different in the second antinomy.

[95] This is true, for example, of the way Baumgarten talks of composition. He says that a "composite thing" in the narrow sense is a "whole of parts outside of parts," i.e., a thing that has parts that are substances. He goes on to say that a simple thing—i.e., not a composite thing—is a monad, that every substance is a monad, and that a composite thing is not a monad but is "composed of monads" (see *Metaphysics*, §§224, 225, 230, and 235). But he clearly does not mean that a composite thing is "composed" of monads in any spatial sense. It is not as though a bunch of monads are thrown together to somehow form a composite thing. This is clear from what he goes on to note about monads: "A monad is not extended and does not fill space. But a whole of monads is extended" (§242). He goes on to clarify that space arises as a result of certain relations among (non-spatial) monads, and that furthermore, in every composite thing in the narrow sense, there is space (§§238, 239, and 241) (Watkins 2009).

on the existence of simples. Likewise, we could say that the "conditioned" thing, the "composite substance" is ultimately conditioned by a simple thing. The narrow interpretation leaves us with the conditional conclusion that *if* matter is a substantial composite, it must be "composed" of simples in the sense that its existence ultimately depends on that of monads.[96] And indeed, the second antinomy is about the constitution of the world as a whole; this interpretation states that for us to experience matter as a composite *substance*, the world as a whole must ultimately contain immaterial simples as its fundamental constituents.

I have been speaking as though these fundamental constituents must be interpreted as Leibnizian monads—but indeed, the conclusion is just as compatible with Platonic ideas—which is a point in favor of my interpretation, since Kant aligns the thesis with Plato. The operating idea in the thesis argument is that our phenomenal experience of sensible objects depends on the existence of simples. This is true for Plato, since for him, the intelligible world, the world of forms, is what explains the phenomenal aspects of our experience. Of course, for Plato, a phenomenal object is explained by its "participation" in the forms—it is not "composed" of forms. But if we take "composed" in the sense for which I have argued, we see that the phenomenal objects depend for their existence on the existence of the forms. Furthermore, these forms are "simple" insofar as they are unified and singular.[97]

---

[96] One possible objection to this interpretation is that Kant seems to state that the thesis argument is not about monads. I deal with this objection below in section 4.

[97] In the *Phaedo*, for instance, Plato calls the forms "non-composite" (78c). He says in the *Republic* that a form is "not many, but one" (476a). We know (from his remarks at A 316/B 372) that Kant was familiar with Brucker's remarks on Plato and Brucker summarizes Plato's views by saying that for Plato "The

Now we are in a position to clarify the claims of the empiricist. The conclusion of the antithesis states that "no composite thing in the world consists of simple parts, and nowhere in it does there exist anything simple" (A 435/B 463). We have seen that the rationalist does not presuppose that composite substances are material. The empiricist does not either, but it will turn out that for the empiricist, a "composite substance" is necessarily an extended or material substance. Just as the rationalist conclusion hinges on a Leibnizian relational view of space, the empirical conclusion hinges on the Newtonian view that space is logically prior to the objects it contains.

As we have seen, Leibniz thought space is grounded in relations among monads or immaterial simples. This is why for Leibniz space and time are "only determinations or relations of things," (A 23/B 37), which is to say that the existence of space depends on the existence of monads.[98] For the Newtonian, on the other hand, objects have relationships to each other *in virtue* of their spatiotemporal placements. Space, for a Newtonian, is an absolute container, an "actual entity," as Kant says, that material objects exist within (ibid.). For Newton, space is logically prior to and independent of substances.

Given the Newtonian views about the relationship between space and physical objects, we can see how the empiricist will argue for the idea that a composite whole does not consist of simple parts. A composite thing, for the empiricist, is just something

---

human intellect is employed … upon things which it comprehends by itself, and which are in their nature simple and invariable" (Enfield 1791, 127).

[98] For Leibniz's argument that space is ideal, see (Leibniz and Clarke 2000, Leibniz's Fifth Letter §47).

composed of parts. But something can only have substantial parts *in space*, otherwise, there would be no external relations between the parts and hence nothing *composite*. But since we are discussing something composite by definition, those parts must exist in space, and since they exist in space, "there must exist as many parts of space as there are parts of the composite thing occupying it" (A 435/B 463). Matter, as a substance, is dependent on space for its existence and thereby takes on whatever mathematical properties are inherent in space, including infinite divisibility. If matter is infinitely divisible, then we will never reach a component that cannot be further divided into parts. The empiricist concludes that a composite substance, i.e., matter, does not consist of simple parts.

Notice that this conclusion is perfectly compatible with the empiricist believing in the existence of physical atoms. Even if an atom cannot be actually divided, it does not count as simple, since anything spatial is not simple. In this sense, both the rationalist and the empiricist agree that a simple must be immaterial. The difference in their arguments is that while the rationalist thinks that the existence of any composite substances ultimately depends on the existence of immaterial simples, the empiricist thinks that the existence of any composite substances ultimately depends on space. The empiricist's conclusion is also compatible with the contention that ideas might have a certain type of simplicity. It is just that for the empiricist, ideas are not fundamental constituents of the universe but rather contingent products of human imagination (A 853/B 881).

Likewise, the rationalist can agree that matter is dependent on space for its existence and that it thereby takes on the mathematical properties of space, including

infinite divisibility. Leibniz for one agrees with this. But for Leibniz, and for the generic rationalist whose argument Kant represents in the thesis, this is a reason in favor of the conclusion that ultimately something simple must ground material objects; they think space alone cannot adequately explain the existence of material objects. Recall the rationalist's reasoning : if you remove "in thought" all of the parts of matter, either 1) you will have nothing left in thought, i.e., the "reality of matter disappears into nothing" or 2) the reality of matter disappears into that which is no longer matter, namely, the simple (A 413/B 413 and A 434/B 462). In other words, the rationalist believes that if matter is ultimately dependent on nothing but space, it would be impossible to explain its substantial reality. The alternative is to say that the reality of matter is ultimately conditioned by something non-spatial, and hence non-material. Thus, for the rationalist, the simple that grounds the reality of matter is immaterial or intelligible; matter gets its reality not in virtue of anything in our spatiotemporal experience but in virtue of the way we necessarily conceive it. This is to say that matter is grounded in something that is only intelligible and cannot be perceived spatiotemporally.

In summary, the rationalist and empiricist offer opposite and mutually exclusive explanations for the existence of substantial wholes. The rationalist argues that the existence of substantial wholes depends on, and can only be explained by, the existence of immaterial simples, e.g., monads or ideas. The empiricist, on the other hand, argues that the existence of substantial wholes, specifically material objects, depends on their existence in space, which implies that no simple entities exist. Next I will turn to Kant's resolution to this antinomy.

3. Kant's Resolution to the Second Antinomy

As we have seen, reason supports both the conclusion that everything in the world is either simple or made up of simples *and* the conclusion that nothing in the world is made from simples. The law of non-contradiction implies that both conclusions cannot be true, and the law of the excluded middle implies one or the other must be true. Since neither reason nor experience can adjudicate the debate, we could choose to remain skeptical. On the skeptical view, though we assume that reality itself is non-contradictory, we accept that reason cannot answer all of our philosophical questions. Kant does not accept any such treatment of the antinomy, since he thinks reason should be able to answer its own questions (A 476/B 504).

Kant argues that transcendental idealism is the key to resolving the antinomy (A 491/B 519). The cosmological contradictions arise, according to Kant, only if we assume that appearances are things in themselves. In regards to the second antinomy, transcendental idealism is supposed to show that both conclusions are false: substantial wholes are not constituted by simples *or* of infinitely divisible parts. We can *think* of substantial wholes *as if* they were divisible to infinity, but doing so does not amount to the claim that they are. Kant's solution, in other words, distinguishes between a prescriptive rule—a "regulative principle"—and a description of reality. We can continue to divide substantial wholes infinitely without ever expecting that we will reach an "unconditioned" part. But that does not mean that substantial wholes *are* infinitely divisible. In this section I explain what this distinction amounts to and how transcendental idealism resolves the antinomy.

First, it will be helpful to understand Kant's general strategy in resolving the cosmological antinomies. He says each antinomy is based on a sophistical argument:

> [1] If the conditioned is given, then the whole series of all conditions for it is also given.
> [2] Now objects of the senses are given as conditioned;
> Consequently, etc. (A 497/B 525)

Reminiscent of the second and third paralogisms, Kant does not finish the argument. But we can assume that the conclusion is meant to say, "consequently, the whole series of all conditions for objects of the senses is also given." In regards to the topic of the second antinomy, this means that when the substantial whole is given we assume that the whole series of all conditions for it is given. In other words, if the "series of conditions" for a substantial whole is simple parts, we can assume that all of those simple parts, being conditions of the substantial whole's existence, are also given. Likewise, if the "series of conditions" for a substantial whole is infinitely divisible parts, we can assume that all of those infinitely divisible parts are given. Both inferences are based on the more general principle stated in the major premise that when a conditioned is given, the whole series of conditions is also given.

Importantly, Kant characterizes the above syllogism in the same way he characterized the paralogistic arguments: as a *sophisma figurae dictionis*, i.e., a fallacy of equivocation (A 500/B 528). He continues the analogy between the antinomies and the paralogisms by noting that the major premise takes the conditioned in the "transcendental signification" of a pure category and the minor premise takes it in an "empirical signification" of a concept of the understanding, which was his analysis, recall, of the paralogisms. Given this, we have good reason to believe that we should analyze the

antinomies in the same way I argued we should analyze the paralogisms (chapter 3). We should expect that each argument goes wrong because of a lack of transcendental reflection, i.e., a failure to modify one's judgments according to the particular way of approaching the object, whether through sensibility or the understanding. Indeed, as I will show, this is exactly how we can characterize the mistakes of the second antinomy.

Kant says that the major premise of the above argument—that "if the conditioned is given, then the whole series of all conditions for it is also given"—is true if we qualify it to say that "if the conditioned is given, then through it a regress in the series of conditions for it is **given** to us **as a problem**" (A498/B 526, Kant's bold). What Kant means here is that the concept of a conditioned necessarily implies that it has a condition. But the qualification that the conditions are given to us "as a problem" indicates that when dealing with appearances, we cannot necessarily assume that because every conditioned thing has a conditioned, that all the conditions are *actually given*. Rather, the only thing we can assume is that the "series" is "given to us as a problem," which is to say that we can continue to find conditions. For the whole series to be "actually given to us" would mean that all of the conditions are presented to us simultaneously. But the very nature of the object we are speaking of, a completely divided whole, cannot be given to us that way, since our experience is necessarily temporal. We cannot experience all the conditions at once, since they are necessarily presented to us successively through time.

Given the above analysis, we can see why Kant goes on to say that the claim that "if the conditioned is given, then the whole series of all conditions for it is also given" is the product of representing things "**as they are** without paying attention to whether and

how we might achieve acquaintance with them" (A 498/B 526-7, Kant's bold). We will recall from chapter 3 that to ignore the distinctive contribution of either sensibility or our intellect is to fail to transcendentally reflect. Thus, the antinomial conflict is the result of a lack of transcendental reflection.

This last point will become clearer if we examine Kant's resolution of the second antinomy in particular. Kant's assessment is that a substantial whole is neither constituted by simples nor of infinitely divisible parts. The law of the excluded middle, he implies, only applies under the assumption that a substantial whole is a thing in itself (A 501/B 529). But we should not make such an assumption, he continues, since doing so leads to the contradictory conclusion that a substantial whole must be composed of simples *and* that it must be composed of infinitely divisible parts. The very contradiction, Kant is trying to show, is *proof* that the initial assumption is wrong. Thus, a substantial whole is not a thing in itself (A 506-7/B 534-5). Transcendental idealism resolves the antinomy and the antinomy is indirect proof of transcendental idealism. Indeed, according to Kant, transcendental idealism is true *because* it resolves the antinomy.

We should be clear on what it means to say that a substantial whole is neither constituted of simples nor of infinitely divisible parts. Again, we should not take Kant to be making a point about the existence or non-existence of physical atoms. "Simple" entities, as I argued in the last section, refer to immaterial entities such as Leibnizian monads or Platonic ideas. The empiricist is right, Kant thinks, to assume that we will never reach something "absolutely unconditioned" when we are dividing a substantial

whole. Indeed, the "regulative principle" in regards to this antinomy is that we should act as though we can keep on dividing a substantial whole without ever reaching a simple entity. Kant's point is that this truth does not entail what the empiricist thinks: namely, that a substantial whole *thereby* consists of infinitely many parts, or, in the language I used above, that it thereby can exist without relying in some way on the existence of simples. We might ask what the difference is between the conclusion that a whole consists of infinitely many parts and the assumption that a whole can be divided infinitely. Kant's point is that the difference is significant: the conclusion that a substantial whole consists of infinitely many parts *rules out* other conclusions, while the prescriptive rule does not.

Importantly, to say that we can keep dividing a substantial whole is not to disprove the existence of immaterial simples, i.e., it does not show that Leibnizian monads or Platonic ideals *are not* what ultimately ground or explain the existence of substantial wholes. In other words, the fact that we can infinitely divide matter does not prove that the existence of matter is completely and sufficiently explained by the fact that it exists in space. On the other hand, to say that we can infinitely divide matter does not prove that we *must refer* to immaterial simples (monads or ideas) in order to explain the existence of a substantial whole, which was the reasoning of the rationalist. So the regulative principle that we can infinitely divide matter is different from the empiricist's conclusion that a substantial whole *is* infinitely divisible insofar as the first does not entail any substantial metaphysical conclusions (pun intended).

Recall that the conclusions of the second antinomy were proposed to answer the question of what ultimately grounds or explains the existence of substantial wholes. Kant's resolution focuses on proving the rationalist's and empiricist's answers to be false. But what about a positive answer from Kant? If neither the existence of simples nor the existence of infinitely divisible parts explains the existence of substantial wholes or matter, then what does, according to transcendental idealism? The answer is more easily seen if we take a deeper look at how the concepts of comparison and the activity of transcendental reflection work in regards to the cosmological idea.

The second antinomy is aligned with the category heading of "quality," which indicates that the relevant concepts of comparison are agreement and opposition. Agreement and opposition, as I detailed in the last chapter, ground the logical forms of affirmative and negative judgments respectively (see chapter 3, §3d). This is as we would expect: the dueling conclusions of the second antinomy are that the "world *is* composed of simples" and the "world *is not* composed of simples." The judgments presuppose a comparison of a composite substance with the concept of simplicity via the concepts of agreement and opposition. Now, of course, the rationalist and the empiricist both commit a transcendental amphiboly, which as we saw, is a confusion of the object of pure understanding with the appearance. This has motivated their respective arguments.

The rationalists, on the one hand, think that simples must exist to ground the existence of substantial wholes because otherwise, nothing would remain if "all composition of matter were removed in thought" (A 525/B 553). We can see here, however, that this conclusion only follows so long as the rationalist relies only on the

logical condition for agreement and opposition, which, we will recall, states that a property agrees with another (and hence can ground an affirmative judgment) if both can be attributed to the same object without resulting in a logical contradiction . Recall that the rationalists argued for the thesis via a *reductio*, meaning that they think they have shown composite substances are *opposed* to composite parts, yielding the conclusion that composite substances are composed of simples.

Additionally, the rationalist uses a purely conceptual notion of substance in referring to a "composite substance." That is to say, they refer to substance insofar as we must conceive of it as something absolutely inner. But as Kant reminds us, "with that which is called substance **in appearance** things are not as they would be with a thing in itself which one thought through pure concepts of the understanding. The former is not an absolute subject, but only a persisting image of sensibility, and it is nothing in intuition, in which nothing unconditioned is to be encountered anywhere"(A 525/B 553). Hence, in arguing for the thesis, the rationalist has been led astray by what we might think of as a double transcendental amphiboly, first with their notion of substance and second with their unqualified use of the logical conditions for agreement and opposition.

Hence, just as I showed in the last chapter in regards to the paralogisms, transcendental reflection would reveal to the rationalists that they need to qualify their claims. First, they would see that the "composite substance" they refer to in the thesis argument refers to composite substances only insofar as we abstract from spatiotemporal conditions. This means that they are not entitled to conclude anything about "every composite substance in the world," as the thesis argument does (A 434/B 462). Second,

the rationalists would recognize that while we must think of those composite substances as simples insofar as we consider them *logically*, this entails nothing about the way those substances exist spatiotemporally. Hence, they are not entitled to conclude that "nothing exists *anywhere* except the simple," only that nothing exists except the simple insofar as we consider things logically.

A similar argument applies to the empiricist's reasoning in the antithesis. The empiricists argue that the world *is not* composed of simples, which is a judgment that presupposes a comparison of a composite substance with the concept of simplicity via the concepts of agreement and opposition. They too argued by a *reductio,* concluding that the concept of simplicity was opposed to the concept of a composite substance. Just as the rationalist relied on merely the logical conditions for agreement and opposition and think that it yields judgments true of *any* object, the empiricist relies only on the spatiotemporal conditions for agreement and opposition and takes them to yield judgments true of *any* object. And notice, one cannot attribute the property of simplicity to a composite substance insofar as one considers it spatially, since a substance in space will always have at least spatial parts. Hence, the empiricist too must transcendentally reflect on their claims to recognize that they are not entitled to rule out the conclusion that logically and conceptually, simple parts are necessary for composite substances.

Kant's positive answer then, to the question of what ultimately explains the existence of substantial wholes, is that we can only explain their existence *as appearances* by referencing both our sensibility and our understanding. First, substantial wholes are always given to us in space, insofar as we are passive in receiving our

intuitions of them and insofar as they are always spatial. But—and this is the role of the understanding—these substantial wholes can be *substantial* only if we conceive of them as being absolute subjects, which is to say that we must conceive of them as "composed" of simples.

Kant's explanation for the existence of substantial wholes, notice, is not about how metaphysical objects exist. Rather, it is an explanation *about us*. More than anything, it reveals to us truths about our own capacities. Indeed, Kant's argument in the Antinomy is an instance of his "Copernican Revolution" that I discussed in chapter 1. I said there that Kant aims to reorient us in our metaphysical questions. Instead of trying to answer what matter is really composed of, for example, Kant's experimental method shifts the focus to what we are interested in as humans when we ask such a question. In particular, I argued that Kant's method will allow us to pursue our scientific investigations without thinking we need to rely on supersensible entities to do such, and that doing so will give genuine significance to the idea of the cosmos as a whole in a way compatible with our moral agency.

What are we really interested in when we ask what the universe is composed of? Undoubtedly, we care for scientific reasons. But more importantly, the composition of the universe matters to us, Kant thinks, because it appears to have profound implications for us as moral agents. Namely, it reveals whether or not we can exist as the type of entities that can claim responsibility for our own actions, insofar as we might think that simplicity is a necessary aspect of that type of entity. Indeed, textual evidence indicates

that *our* possible constitution is what Kant thought we really care about when we ask the question of the second antinomy. It is to this topic that I now turn.

4. Our Interest in the Second Antinomy

Kant explicitly connects the second antinomy with the soul in at least three passages, all of which occur in the section of the Antinomy titled "On the interest of reason in these conflicts" (A462/B 490 ff.). In each of these passages, Kant lists the four topics with which the Antinomy concerns itself and in each, he connects the second antinomy with the soul (I have emended each with a note indicating the antinomy to which Kant alludes):

> [1] The questions whether the world has a beginning and its extension in space a boundary [first antinomy]; whether there is anywhere, **perhaps in my thinking self, an indivisible and indestructible unity, or whether there is nothing but that which is divisible and perishable** [second antinomy]; whether my actions are free or, like those of other beings, controlled by the strings of nature and fate [third antinomy]; whether, finally, there is a supreme cause of the world, or whether natural things and their order constitute the ultimate object [fourth antinomy] ... (A463/B 491, my emphasis).

He does this again a few paragraphs later, this time focusing on the "thesis" arguments of the antinomies:

> [2] That the world has a beginning [first antinomy], that **my thinking self is of a simple and therefore incorruptible nature** [second antinomy], that this self is likewise free and elevated above natural compulsion in its voluntary actions [third antinomy], and finally that the whole order of things constituting the world descends from an original being [fourth antinomy] … (A 466/B 294, my emphasis).

He makes the connection again a few paragraphs later, this time focusing on the "antithesis" arguments of the antinomies (in a slightly different order):

> [3] If there is no original being different from our world [fourth antinomy], if the world is without a beginning and also without an author [first antinomy], if our will is not free [third antinomy] and **our soul is of the same divisibility and corruptibility as matter** [second antinomy] … (A 468/B 496, my emphasis).

The context in which these passages appear is of great importance. As the title of the section suggests, Kant is discussing here the *interest* we take in answering cosmological questions. At every turn, Kant suggests that our desire to answer the question of what fundamentally constitutes the universe is motivated less by a theoretical or scientific interest than by the profound implications an answer has for the way we are entitled to think of our*selves*. Indeed, Kant says that answering such cosmological questions imbues philosophy with a dignity unequal to other sciences since such questions represent the "highest and most important ends of humanity" as he continues to say after passage [1]. We have a "practical interest" in knowing whether or not we could possibly be simple and therefore incorruptible—since this belief, as passage [2] continues to say, is one of the "cornerstones of morality and religion." Likewise, if it turns out that the soul is of the "same divisibility and corruptibility as matter, then **moral** ideas and principles lose all validity," as he continues to say after passage [3] (Kant's emphasis).

Kant clearly thought that the question of the composition of the soul is part of what we care about when addressing the question of the second antinomy. He unambiguously connects the two. So why have commentators not followed suit? The most obvious reason is that Kant himself says that the thesis—the argument for the existence of simples—does not apply to Leibnizian monads (A 442/B 470), which are essentially mental simples or souls. Kant's reasoning for this is that a Leibnizian monad is given *as a simple* (i.e., in self-consciousness) and not a composite substance, which is what is properly under discussion in the antinomy. Commentators have taken this as evidence that Kant was not thinking of the composition of the soul when he constructed

the second antinomy. Karl Ameriks says, for example, "the soul is not given as complex and thus the topic of its nature is not settled by the discussion" (Ameriks, 51). Although he acknowledges that Kant does go on to connect the second antinomy with the soul (by referring to what I have labeled as passage [2] above), he speculates that this is perhaps the "sign of an earlier plan for the Dialectic that wasn't realized" (Ameriks, 80).[99]

This objection is easily dealt with by recognizing that Kant distinguishes between two ways we can view the soul—what I will call 1) the perspective of self-consciousness and 2) the "external" perspective. When considered from the latter, Kant is clear that the soul is indeed given as a composite, not a simple. Hence, Kant's contention that the thesis does not apply to Leibnizian monads properly construed is to point out that a monad is only properly considered from the perspective of self-consciousness: "the proper signification of the word **monas** (in Leibniz's usage) refers only to the simple given **immediately** as simple substance" (A 442/B 470, Kant's emphasis). When considered from the external perspective, however, we do not immediately perceive the soul as simple, but must argue for its simplicity, given that a thinking thing *appears*, externally, as a composite thing. In the next section I examine this external perspective in more detail.

[99] Grier likewise agrees that the reason why Kant says that the thesis does not apply to Leibnizian monads is that Kant is dealing here with things given as composites. She correctly uses this as evidence that the thesis leaves room for interpreting substance as non-extended, despite the fact that Kant says it does not refer to Leibnizian monads. But again, she immediately rejects the idea that Kant intended the antinomy to apply to the soul (2001, 197).

5. The External Perspective

Kant says in the Paralogisms that the representation "I" "serves to distinguish two kinds of objects through the nature of our power of representation. I as thinking, am an object of inner sense, and am called 'soul.' That which is an object of outer sense is called 'body'" (A342/B 400).[100] Indeed, the self shows up in our experience in a way that is more obvious than the purely conceptual way the rationalist considers it. While it is true that we have good theoretical reason to be interested in what explains the possibility of thought, we are *humans*, who not only think, but have bodies, act, and hold ourselves and others responsible for those actions. Thus, the "soul," insofar as it refers to a human self, clearly shows up in experience.

Kant says exactly this in the second antinomy. At the end of his discussion of the antithesis, he applies the general discussion of composition to the particular question of the composition of the soul. The passage is worth quoting in whole:

> Thus self-consciousness is such that because the subject that thinks is simultaneously its own object, it cannot divide itself (though it can divide the determinations inhering in it); for in regard to its own self every object is absolute unity. Nonetheless, if this subject is considered **externally**, as an object of intuition, then it would indeed exhibit composition in its own appearance. This is the way in which it must be considered, however, if one wants to know whether or not there is in it a manifold of elements **external to** one another (A 443/B 471, Kant's emphasis).

To paraphrase, from the perspective of self-consciousness, one cannot help but think of oneself as a simple, indivisible unity, since the representation "I" is necessarily a non-

[100] Also see *Anthropology from a Pragmatic Point of View*, where Kant says that "now here the "I" appears to us to be double (which would be contradictory): 1) the "I" as *subject* of thinking (in logic), which means pure apperception (there merely reflecting "I"), and of which there is nothing more to say except that it is a very simple idea; 2) the "I" as *object* of perception, therefore of inner sense, which contains a manifold of determination that make an inner *experience* possible" (7:135fn) (Kant 2007, 246).

composite representation. That is, when we use the word "I" to represent ourselves as thinkers, we necessarily take it to represent something that is a simple, indivisible, "absolute unity." From another perspective, however, from which we consider the soul as an "object of intuition," the thing that thinks appears not as a simple thing but as a composite thing—a thing with parts, e.g., a body or a human being. I will call this the "external perspective." From the external perspective, instead of viewing the soul from the consciousness we have of ourselves as thinkers, we view a soul—a thinking thing—from a third-personal stance, from which it is not given to us as a simple thing but as a composite whole.

From the external perspective, the rationalist and empiricist will both have arguments concerning the possible constitution of the soul. While silent on what the soul is like merely from the phrase "I think," the empiricist now has something to say. The rationalist too can inhabit the external perspective. Since from this perspective we do not *immediately* perceive the soul as simple, the rationalist can no longer argue directly from the necessarily simple representation "I" to the simplicity of the soul. The rationalist will claim instead that we must rely not on our sensibility but on our understanding to cognize what the soul ultimately *is*, despite its sensible appearance (A 853-4/B 881-2).

This means that from the external perspective we are led to reason about the soul in a way that leads to a contradiction: that it must be a simple thing (or something "composed" of simples) *and* a composite thing. Notice the parallel with the second antinomy here. The first alternative will lead the rationalist to argue that the soul is immaterial, incorruptible, and immortal; the second will lead the empiricist to argue that

it is material, corruptible, and mortal.[101] The composition of the soul, when considered from the external perspective, is a two-sided illusion.

I am suggesting that reasoning about the soul generates a distinct antinomy, not articulated by Kant, which parallels the second antinomy. This is to say more than that the second antinomy merely has implications for how we think of the soul. It surely does: if we think the soul must be a simple entity for it to be immortal, imperishable, and incorruptible, then the antithesis conclusion that there exists nothing simple would rule out this existence of this type of soul. Likewise, while the thesis argument would not prove the existence of such a soul, it would leave room for it. I am claiming, however, that we can go further and generate an antinomy that directly concerns the constitution of the soul, insofar as we view it from the external perspective.

Evidence for this is found in the passages above. Admittedly, the first passage indicates the constitution of the universe merely has implications for the constitution of the soul. Kant says this explicitly: the question, he says, is "whether there is anywhere, *perhaps* in my thinking self, an indivisible and indestructible unity, or whether there is nothing but that which is divisible and perishable" (A 463/B 491, my emphasis). But notice the stronger language of the second passage, where the concern is whether "my

[101] Early modern empiricists would not be classified as such "materialists." Locke for instance, recommended we remain agnostic about the existence of an immortal soul and argued that we do not need to assume an immaterial soul for the sake of religion. Berkeley, likewise, held the soul to be a spiritual substance. Hume certainly did not think the soul was immortal, but he does not argue for the claim the way I suggest in this chapter (see Hume's "On the Immortality of the Soul"). Hence, we must take Kant to have in mind the Ancient empiricists, e.g., Democritus and Epicurus. Kant gives us a hint that this is what he means when he aligns the antithesis arguments in the Antinomy with Epicurus (A 471/B 299). As I have mentioned before, Kant was familiar with Brucker's *History of Philosophy*. In that text, Brucker clearly identifies Epicurus as a materialist in regards to the soul: "The soul is a subtle corporeal substance, composed of the finest atoms" (Enfield 1791, 273). He also clearly identifies Democritus as a materialist: "the soul, or principle of animal life and motion is the result of a combination of round or fiery particles" (ibid., 249).

thinking self is of a simple and therefore incorruptible nature" (A 466/B 293) and likewise of the third, where the concern is whether "our soul is of the same divisibility and corruptibility as matter" (A 468/B 496). The last two passages, in other words, show a direct concern with the constitution of the soul, not just about what the constitution of the universe implies about it.

One might think, along with Ameriks, that such passages are simply a remnant of an earlier plan for the Antinomy. Indeed, Kant's unpublished notes indicate that the constitution of the soul was a topic originally included in the Antinomy before Kant decided the soul deserved its own chapter (see especially R 4757). But this only supports my claim, which is that there is a separate, distinct antinomy of the soul, not articulated by Kant in the first *Critique* but suggested by the structure of the second antinomy. We can only surmise why Kant himself did not follow through with such a plan. One possibility is that he did not find the threat of materialism a real one. Such a threat is real for us, however. Hence, it is fruitful to examine what an antinomy of the soul might look like. I offer one possibility in the next section.

6. The Antinomy of the Soul

Using the second antinomy as a model, we can construct what an antinomy of the soul might look like. Recall that an antinomy in general begins by asking what ultimately grounds some aspect of our appearance. Accordingly, this antinomy will begin by asking what has to be true of a thinking thing—what ultimately must ground it—to explain its appearance as a composite whole such as a human being, body, or brain. The arguments will give conflicting explanations of how this can occur.

The rationalist, arguing for the thesis, will argue that we can ultimately explain the soul's appearance as a composite whole only if it ultimately composed of (in the sense I argue for in section 2), or is itself, a simple, immaterial thing. This is tantamount to saying that there would no physical substances in the world that appeared to think or instantiate thinking without there also being some kind of mental, immaterial substance grounding that appearance. Hence, the thesis argument could represent a range of diverse views, including Cartesian dualism, Leibnizian monadism, or Platonic idealism.

The empiricist, on the other hand, arguing for the antithesis, will argue that we can explain the soul's appearance as a composite whole only insofar as we consider it a spatial entity. This is tantamount to saying that there would be no physical substance in the world that appeared to think or instantiate thinking without it also being spatial. Furthermore, the empiricist would argue that we do not need to reference immaterial entities like monads, ideas, or mental substances in order to explain substantial wholes that think. Hence, the antithesis argument could also represent a range of diverse views, including physicalism, identity theory, eliminativism, or reductive materialism.

a. The Thesis: The Soul Must be Simple

The thesis of the second antinomy argues for the claim that "every composite substance in the world consists of simple parts, and nothing exists anywhere except the simple or what is composed of simples" (A 434/B 462). Applied to the soul, the rationalist would argue that a soul, or thinking thing, is ultimately composed of something simple, (again, with the caveat that "composed" here means grounded on or explained by). As I said above, we might say that the soul, from the external perspective,

is something like the brain or the body, i.e., something composed of physical matter. Indeed, in our experience, thinking things always have bodies. Kant also notes, however, that thinking of the soul as "composite" could mean that it is composed of powers or faculties of one and the same substance (B 416fn). Regardless, the rationalist is trying to prove here that a thinking thing must ultimately be composed of or be something simple.

The rationalist's argument, mirroring the second antinomy, would proceed indirectly: assume that a thinking thing is not composed of anything simple. If we "remove all composition" in thought, nothing would be left over—there would be no substance, i.e., no brain or body. More importantly, since we are assuming a thinking thing is not composed of something simple, we are also assuming that there is no subsisting "I," since the representation "I" is necessarily simple. We have thereby eliminated the subject of thought, i.e., the thing that thinks. So, a thinking thing, even if it appears to have parts, must ultimately be or be explained in terms of something simple.

Indeed, the rationalist is going to argue that whatever it is that appears externally as a thinking thing is not *actually* the thing that thinks. This is because according to the rationalist, a thinking thing can only be represented as something that is self-conscious. But notice that here we are engaged in reasoning from the external perspective, not the perspective of self-consciousness (A 347/B 405).[102] The rationalist's insistence that the thinking thing as it appears externally cannot actually be a thinking thing aligns exactly with how Kant says the rationalist will view reality: sensible objects, for the rationalist,

[102] Ultimately, Kant himself agrees that this is the only way to represent a thinking thing. He does not, however, follow the rationalist in concluding that this means that it must *actually be* a simple thing. The empirical view that it is composite, despite our necessary representation of it through self-consciousness, is still an option (see especially A 363-4fn).

are “mere semblance” and truth is to be found only in what we cognize through our understanding (A 854/B 882).

b. The Antithesis: The Soul Cannot be Simple

The antithesis of the second antinomy argues for the claim that “no composite thing in the world consists of simple parts, and nowhere in it does there exist anything simple” (A 435/B 463). Applied to the soul, the empiricist would argue that the soul cannot be a simple thing. Start by assuming that the soul is a simple substance. Everything substantial exists in space, so if the soul is a substance, it too must exist in space. Assume, then, that the soul is a spatial substance. It would then consist of a “manifold of elements,” e.g., spatial parts, and therefore, the presumed simple soul would also have to be composite. This is a contradiction; therefore, the soul is not simple. Furthermore, the empiricist will go on to argue, we could never establish that the soul is simple by any experience or perception for the same reason that we can never establish any substance to be simple. Hence the claim that the soul is simple is merely an idea that has no objective reality. The empiricist’s assumption that the way we represent an object, i.e., through sensibility, represents how that object *is*, entails that no simple thing exists in reality since we cannot sensibly represent it as such.

c. The General Conflict

It is good to take a step back and remind ourselves of the spirit of these arguments. The conflict concerning the soul is a conflict of two broad views: one that thinks that everything real is material and concrete, and thereby accessible to our senses, and one that thinks that material substance is a mere “semblance” of a non-material

reality and that knowledge of this reality ultimately lies in what we can understand through reason (A 854/B 882). These conflicting metaphysical views, when operating under the assumption that the way we cognize an object represents how it really is, lead us to conflicting views about the composition of the soul. But we should not forget that this conflict, according to Kant, is not one of mere metaphysical sparring, it is one that arises naturally for humans (B 354/A 298). This implies that the conflict ought to arise without first taking a position on what is ultimately real.

If we start intuitively from the question of what the soul is, we can see how we might be led naturally to both views. A soul is a thing that thinks and it seems as though, from the nature of the representation "I," that this thing needs to be different from material objects. We can imagine removing our body and still having the representation "I." In some sense, then, we might conclude that the soul must be immaterial and not spatial like other objects. On the other hand, every thinking thing, in our experience, is embodied—and indeed, we naturally take an utterance of the word "I" to refer to a *person*, not a disembodied mind. We might conclude that the "soul" is like other material objects and hence subject to corruption and destruction.

Just as we saw that Kant thinks transcendental idealism will resolve the contradiction of the second antinomy, it will resolve the contradiction inherent in our reasoning about the soul. I offer such a resolution in the next section

7. The Resolution to the Antinomy of the Soul

As Kant indicates in passages [1], [2], and [3] above, simplicity is important because it implies indivisibility, indestructibility, and incorruptibility, and ultimately

immortality. Hence, the soul's antinomy can be seen as a conflict between two more detailed views about the soul: 1) the view that the soul is indivisible, indestructible, incorruptible, imperishable, and immortal, and 2) the view that the soul is divisible, destructible, corruptible, perishable, and mortal.

Just as reason supports both the thesis and the antithesis of the second antinomy, it supports both the conclusion that the soul is simple and immaterial and that it is composite and material. Using Kant's resolution to the second antinomy as our model, we can construct a possible Kantian resolution to the antinomy of the soul. The general conclusion would be that the soul, when viewed as a composite thing, is neither simple nor composite and that the mutually exclusive contrast is only the result of the false assumption that the soul as a composite whole is a thing in itself. Once rid of this assumption, we can see that we cannot claim to have knowledge of the soul as a material object or as an immaterial object. Indeed, in regards to the second antinomy, Kant says that both the thesis and antithesis are false (A 528/B556). If we follow him in regards to the antinomy of the soul we would say that the soul as a composite whole is *non*-material (see also A 503/B 531).

Just as with the resolution to the second antinomy, we might scoff here: aren't souls either material or not? Apart from the concerns about the law of the excluded middle, what does it mean to say that the soul is "non-material"? What kind of thing *is* the soul according to Kant? Recall that we are speaking of souls *insofar* as they are given as composite wholes—i.e., thinking things insofar as those things appear to us as whole entities with parts, e.g., human beings, bodies, or brains.

To say that the soul is non-material, I argue here, is to offer a prescription rather than a description. A Kantian resolution to the antinomy of the soul will shift our focus away from what constitutes the soul to how we can legitimately treat it given our human commitments. The prescription is that we may treat the soul either as a material thing or as an immaterial thing, but that neither of these treatments has claim to the truth in a way that rules the other out. Thus, neither materialism nor immaterialism is a threat to our moral or scientific commitments. It is wrong to say either that the soul is simple and hence indivisible, indestructible, incorruptible, imperishable, and immortal, or that it is composite and hence divisible, destructible, corruptible, perishable, and mortal.

Kant's resolution to the second antinomy replaced a possible description of the universe with a regulative rule. In that case, the regulative rule was in favor of the antithesis argument; it told us that though we should stop short of saying that matter is infinitely divisible, we can act as though it is. In regards to the antinomy of the soul, however, it becomes clear that neither the thesis nor the antithesis can be so favored. This is because of the special nature of the composite whole that thinks; it grounds both scientific *and* (what we might think of as) spiritual concerns.

Scientific concerns include neurological effects on one's identity and what Kant would call empirical psychology—a study of how humans do in fact act in certain situations. Spiritual concerns, in comparison, involve any moral or religious commitments we have regarding ourselves and others, including whether we can be morally responsible and whether we can survive our body's death. Given this range of commitments, we can expect there to be two general regulative principles that arise in

connection with the antinomy of the soul—one for each side of the argument. The first, which we might think of as a spiritual regulative rule, is that we can act as though the soul is a simple, immaterial substance. The second principle, which we might think of as an empirical regulative rule, is that we can act as though the self is inseparable from the body or brain. I will argue in the next chapter that the empirical regulative rule presupposes at least a minimal version of the spiritual one.

To treat the soul as a material entity is to suppose that we can empirically investigate the self insofar as it is a physical entity without ever surmising that we will need to reference metaphysical entities like monads, ideas, or mental substances in order to do so. We should not conclude from this, however, that the mind *is* the brain or body. Any materialist view of the soul cannot claim to exhaustively capture what the soul is; the materialist stance is legitimate only as a rule that regulates our scientific investigations, not as a description of ultimate reality. Thus, an empiricist can never legitimately conclude to have conclusively ruled out that the soul is simple, that is, that it exists in itself, without reference to anything external. While empiricists can make legitimate claims about the self from sensibility, they should not pretend that our sensible knowledge of it does not exhaust it; there is still legitimate room for one to think of oneself abstracted from all of one's spatiotemporal determinations.

Indeed, Kant will argue that we are not able to view ourselves as merely spatiotemporal objects, especially in the realm of morality. This gives rise to another regulative rule associated with the thesis argument of the antinomy of the soul. To treat

the soul as an immaterial entity amounts to a rule that says we should take human beings to be unified entities identical across time. I detail this regulative rule in the next chapter.

THE IDEA OF A PERSON AS A WHOLE: UNITY AND COMPLETENESS IN KANT'S RESOLUTION TO THE THIRD ANTINOMY

CHAPTER 5

In the last two chapters, I considered Kant's arguments concerning what the self is not. Kant thinks it is wrong to think of the self as exclusively material or immaterial. The question now is what Kant's positive view is. The answer is that a "self" for Kant is a material or sensible object the action of which we necessarily conceive of under the idea of freedom, or more broadly, as I will show, under the "idea of completeness." This chapter will address what thinking of ourselves under the idea of freedom or completeness amounts to and will end with a discussion of Kant's theory of how we must think of ourselves in the moral realm.

Kant presents his positive view—that a self is something we necessarily think of under a particular idea—in a section of the first *Critique* called "The Resolution to the Third Antinomy," which is often hailed as one of the most important sections of the book. In the third antinomy, Kant compares the arguments of two seemingly contradictory views: that causality through freedom exists, i.e., that "causality in accordance with the laws of nature is not the only one," and that freedom does not exist but that everything in the world happens "solely in accordance with the laws of nature (A 444-45/B 472-73). Kant's "resolution" is designed to resolve the apparent contradiction between these two views.

Kant does so by arguing that we can think of ourselves in two different ways: as having an empirical character, wherein we think of our choices as causally determined, and as having an intelligible character, wherein we think of ourselves as "free of all

influences of sensibility and determination by appearances" (A 541/B 569). Thus, the resolution supposedly shows how both arguments of the third antinomy can be true. Causality through freedom exists, since we can view ourselves as free, but everything in the world still happens solely in accordance with the laws of nature. Kant says that the former describes an individual as an appearance, the latter as a thing in itself (A 539/B 567).

It is clear then, why most commentators take Kant's resolution to be something akin to compatibilism. The contemporary free will debate is broadly split between compatibilists, who think freedom and determinism are compatible, and incompatibilists, who think they are not. Determinism is the theory that every event is causally necessitated by the laws of nature. Kant is a determinist.[103] As such, it is easy to think that he means to argue that human freedom is compatible with determinism. As most commentators admit, however, Kant's views do not fit comfortably within the confines of the contemporary debate. First, Kant wants to embrace a robust conception of freedom in his moral theory—one that Allison characterizes as "incompatibilitist" (1990, 28). Kant complains that compatibilist conceptions of freedom are "nothing better than the freedom of the turnspit" (C2 5:97).[104] Hence, Kant seems to want to show not just that freedom

[103] See his argument in the Second Analogy that "all alterations occur in accordance with the law of the connection of cause and effect" (B 232 ff.). As we will see, however, this claim needs some qualification. Kant does think the causal principle is true—but as we will see, he does not think determinism explains everything.

[104] Here, Kant probably refers to Leibniz's compatibilist conception of freedom.

and determinism are compatible, but as Wood says, that compatibilism and incompatibilism are compatible.[105]

Kant's resolution, however we characterize it, is normally thought to be unsuccessful and puzzling. Wood, for instance, says that "Kant's solution to the free will problem strikes nearly everyone who has ever studied it as thoroughly unsuccessful, a metaphysical monstrosity that gives us a far-fetched if not downright incoherent account of our moral agency" (1984a, 75). Bennett argues that "noumenal freedom" could only "be consistent with itself and with determinism merely by being vacuous," and goes on to conclude that if Kant's resolution is successful, it is "only because one half of it has no real content" (1967, 194).

In this chapter I will argue that common approaches in the secondary literature misunderstand Kant's resolution to the third antinomy, especially his conception of intelligible character. The common approaches, Wood's and Allison's in particular, err in thinking that Kant's solution aims to show that freedom is *possible*, given the truth of determinism. Undoubtedly, Kant talks this way himself sometimes. But given Kant's overall discussion of how empirical character and intelligible character relate, it is clear that attributing an empirical character to a human being—and hence thinking of him as a causally determined agent—is not possible without first attributing to him an intelligible character, which I will argue is the overarching idea we have of an agent *as* an agent, the idea of a person as a whole. Allison comes close to this type of interpretation, but stops short of it. We will see that developing it fully provides much needed traction to Kant's

[105] See Wood 1984a, 74. Allison agrees with this characterization (1990, 28).

notion of "intelligible character." Instead of it being a weak, arbitrary solution that only a transcendental idealist would accept, Kant's real intention is for it to be a challenge to the determinist or compatibilist. One who desires to attribute an empirical character to a human being must also attribute an intelligible one. Thus, if Kant's resolution is successful, he has shown that human freedom is not just compatible with determinism as applied to human action, but necessary for it.

I proceed in section 1 by summarizing the arguments of the third antinomy. In section 2 I present the commonalities of Wood's and Allison's interpretations of it, which broadly represent the commentary as a whole. I will call these the "standard interpretations." In section 3, I will present my interpretation, which I will call the "explanatory-problem" interpretation. There, we will see that both arguments of the third antinomy present necessary components of explanation that seem incompatible: what I will be calling the completeness requirement and the unification requirement. Section 4 addresses the completeness requirement and human action. Section 5 addresses the completeness requirement and its role in non-human causation. Section 6 addresses how the idea of freedom gains its significance in the moral realm and why Kant thinks he is entitled to the strong claim that we know we are free. I conclude by discussing why Kant thinks the self understood as a moral agent is the proper self.

1. The Third Antinomy and Kant's Resolution

The third antinomy, just like the second, consists of Kant's presentation of two arguments that support opposite claims. Kant's claim in the Antinomy chapter, as we have seen, is that "pure reason" leads to a contradiction when attempting to answer

questions about the universe—that is, it leads us to assert two claims both of which are supported through reason, but cannot, it seems, both be right. The thesis states: "Causality in accordance with the laws of nature is not the only one from which all the appearances of the world can be derived. It is also necessary to assume another causality through freedom in order to explain them" (A 444/B 472). The antithesis states: "There is no freedom, but everything in the world happens solely in accordance with the laws of nature" (A 445/B 473).

The argument for the thesis begins by assuming that determinism is true, or, as Kant puts it, that "everything that happens presupposes a previous state, upon which it follows without exception according to a rule" (ibid.). Then for any given "happening," i.e., any given event, a previous state—a cause—must have already happened. Likewise, that event must too have arisen from something that has already happened, i.e., another cause. It follows from this that there is never a first beginning and thus "*no completeness of the series* on the side of causes" (my emphasis). The very idea of a law of nature, however, assumes that "nothing happens without a cause sufficiently determined *a priori*." Thus, it is contradictory to say that determinism, "when taken in its unlimited universality," is the only type of causality, proving the initial assumption false. The premise that there must be a first beginning and the assumption that "completeness of the series" is necessary in order to "sufficiently determine" a cause *a priori*, is the crux of the argument. In order to sufficiently explain any given instance of cause and effect, I must completely explain it, i.e., provide an explanation that has a *beginning*, the explanation of which does not depend on referring to another cause.

The antithesis argues for the conclusion that "there is no freedom, but everything in the world happens solely in accordance with the laws of nature" (A 445/B 473). The argument proceeds, like the thesis, by an indirect proof. It begins by assuming that causality through freedom exists. This would entail that a causal series would "begin absolutely through spontaneity," i.e., a series would exist the beginning of which is not preceded, and hence not determined by, any constant law. But the "unity of experience" is only possible given the fact that every action presupposes a state that in turn can be causally connected with a preceding event. Thus, it is not the case that causality through freedom exists. The crux of the argument lies in the premise that causality through freedom would violate the *unity* of experience.

Kant argues that in the case of the first two antinomies, the indication of a contradiction proves wrong an assumption of traditional metaphysics, namely, that appearances are things in themselves. So, through a *reductio ad absurdum*, Kant finds indirect proof for the distinction between appearances and things in themselves. Kant resolves the first two antinomies by showing that both the thesis and the antithesis are false. Kant resolves the second two, however, by showing that both the thesis and the antithesis can be true, despite the fact that they rule each other out. We should notice here that Kant is not explicitly concerned with human freedom—what he calls practical freedom—in the third antinomy arguments. Rather, he is concerned with "transcendental freedom," which is a "faculty of absolutely beginning a state" (A 445/B 473), which is a presupposition, he says, of practical freedom. Nevertheless, it appears as though his resolution to the third antinomy focuses solely on reconciling determinism and human

freedom. This is somewhat misleading, as I will show. But for now, we can take a look at the way Kant proposes to resolve the contradiction between freedom and determinism.

Kant says, basically, that we can view ourselves as "intelligible." This is because we are not solely an appearance in the world of sense. We have a faculty—apperception—that is not itself an object of intuition (A 538/B 566). As such, we can view our actions as "intelligible" and their effects as caused both empirically *and* through an "intellectual concept of its causality." This is just to say that we can view the effects of our actions as both determined and free. "Every effective cause must have a character," Kant says, "i.e., a law of its causality, without which it would not be a cause at all" (A 539/B 567). As such, we can describe ourselves as having two types of characters. First, we can attribute to ourselves an empirical character, whereby we view our actions as standing in "constant accordance with natural laws," i.e., as causally determined. Second, we can attribute to ourselves an intelligible character, whereby our actions can be thought of as free.

Kant goes on to say that the agent, under the description of its intelligible character, does "not stand under any conditions of time, " and can "never be known immediately," since it is a thing in itself (A 539-40/B 567-8). Furthermore, an intelligible character's actions have "no connection with appearances as causes" but it can nevertheless be considered the originator of them (A 541/B 569). Furthermore, Kant says that empirical character is the "sensible schema" of intelligible character (A 553/B 581), implying that the former is somehow dependent on the latter. He even goes so far as to say that intelligible character "determines" empirical character (A 551/B 579). This

seems to suggest that one's intelligible character—as a thing in itself—is what causes one's empirical character, a strange claim to make for a philosopher who argued that the concept of "cause" does not apply to things in themselves.

We can see here why readers have trouble with Kant's resolution. For it looks as though Kant is positing a noumenal self that never appears in experience, is unknowable, has no connection with appearances, and yet somehow acts and effects change in the world. Even if this were a plausible picture by itself, it is not clear how it resolves the dilemma posed by the third antinomy. The solution seems to be that even though we must view ourselves as causally determined, we can also view ourselves as free. But as Allison points out, this seems to merely re-describe the original problem on a higher level (1990, 43). The issue remains how to make both types of characters compatible with each other.

2. The Standard Interpretations

In this section, I examine both Allison's and Wood's interpretations of Kant's resolution, particularly how they interpret empirical and intelligible character and the relationship between the two. Although the interpretations are very different from each other, they share certain assumptions regarding what problem they think "intelligible character" is supposed to solve. Likewise, they share assumptions about the relationship between intelligible character and empirical character. As we will see, they both think Kant introduces intelligible character to show that freedom is at least not incompatible with determinism. As I will show in section 3, Kant introduces intelligible character to

do much more than this: he means to show that freedom is a necessary component of any good and complete explanation.

On Wood's reading, Kant makes compatible compatibilism and incompatibilism by attributing to an individual a "timeless agency," wherein a "particular timeless choice of my intelligible character affects the natural world by selecting a certain subset of possible worlds, namely, those including a certain moral history for my empirical character, and determining that the actual world will be drawn from that subset of possibilities" (1984a, 91). In other words, Wood thinks Kant's solution is to project another metaphysical level of sorts, where an individual's actions are free, that in turn determines the phenomenal world, which is in itself completely determined.

Allison interprets Kant's resolution not as a metaphysical positing of two worlds but rather as a manifestation of Kant's more general "two-aspect" view of metaphysics, namely, that empirical and intelligible character are two ways in which we can view ourselves. On Allison's interpretation, empirical character is a description of an agent insofar as that agent is determined not only empirically but also psychologically. Empirical character expresses a type of rational agency, according to Allison, that is enough to account for a compatibilist conception of freedom insofar as it allows an individual's choices to be determined by beliefs, desires, and intentions (1990, 30-31). This compatibilist conception of agency, however, is not enough for Kant, Allison argues, because one's awareness of one's spontaneity through apperception requires one to also attribute to oneself an intelligible character (ibid., 37-38). Intelligible character is

compatible with empirical character because it "regulates… our conception of ourselves as rational agents" (ibid., 45).

Despite their obvious differences, both Wood's and Allison's interpretations of Kant's resolution share common assumptions. First, they tend to present Kant as privileging determinism, and, only as an afterthought, proposing freedom as something that is not at least incompatible with it. Wood, for example, says that "in defending our freedom, Kant concocts a metaphysical theory which, if true, saves our practical freedom despite the fact that our actions are determined by natural causes" (1984a, 84). Likewise, Allison says "the most pressing question … is … whether we can *ever* regard ourselves as free" in the sense that we can act in ways other than the ways we are causally determined to act (1990, 43). Now, as I have mentioned, Kant is a determinist. So we can forgive an interpretation of the resolution that privileges determinism. However, they also assume that for Kant determinism *alone* can explain what the third antinomy is attempting to explain. The reason this is a problem, as we shall see, is because Kant himself actually frames the third antinomy in a way that makes clear that determinism and freedom are both necessary *explanans.* The way that Wood and Allison frame the resolution makes it appear that determinism alone is a necessary explanans and intelligible character is posited to explain how freedom is compatible with that necessity. I argue below that this does not correctly capture what is at stake for Kant: intelligible character is posited as something that explains how two *necessary* but seemingly incompatible explanans are indeed compatible. This makes a subtle but important difference in how we understand the role of intelligible character.

A second but related assumption that Wood and Allison share is that an agent *just has* an empirical character, yet it needs to be shown that it can have an intelligible character. Wood, for instance, says, "if our actions are indeed causally determined by natural events, then they are apparently necessitated by sensuous impulses acting on our empirical character. From this alone it seems to follow that our actions are unfree. How can anything that might be true about our actions from another standpoint render these same actions free?" (1984, 85). Allison goes so far as to say that empirical character alone "is capable of providing the basis for a rich and potentially attractive form of compatibilism…. that would "leave 'elbow room' for freedom in a deterministic… universe" (1990, 34). Now, both scholars go on to admit that for Kant empirical character is in some way dependent on intelligible character. Wood, for example, notes that Kant thinks that empirical character is the "sensible schema" of intelligible character. He goes on to say that "Kant's theory apparently holds that because appearances are not things in themselves, nature is not the complete and self-sufficient cause of events, at least not of human actions" (1984, 87). Likewise, Allison argues correctly that for Kant "empirical character is… seen as of itself insufficient to determine the will…. The missing ingredient is the spontaneity of the agent, the act of taking as or self-determination" (1990, 39).

But while both scholars admit that empirical character is somehow dependent on intelligible character, neither, I will argue, fully develops what this means for Kant. As I will show, Kant means more than just that *for him* intelligible character is necessary to attribute an empirical one. Rather, he means to argue that for anyone it is necessary;

empirical character cannot be attributed to an agent on any view without the overarching idea of a person as a whole, which is, I will argue, just what intelligible character amounts to.

3. The "Explanatory-Problem" Interpretation

The crucial point that commentators seem to forget when interpreting the third antinomy is *why* a transcendental antinomy arises to begin with. Kant is very clear that one arises out of the attempt to explain an appearance. While the Paralogisms and the Ideal inquire into an object of pure reason, the Antinomy inquires into a supposed object that is at least partly given to us within our experience, the universe as a whole. Each antinomy begins with a particular and given aspect of our experience and asks what must be true about the universe as a whole, metaphysically speaking, for us to experience that given aspect. The Antinomy chapter covers four such aspects of the universe: 1) space and time, 2) matter, 3) the fact that certain events, objects, or states of affairs arise or come about, and 4) the fact that some things are dependent on other things (A 415/B 443). Here we are concerned with the third aspect. The purpose of the third antinomy is to answer the question of what grounds or explains the fact that events, objects, or states of affairs arise or come about. This question is connected, of course, with the concept of causality (A 414/B 441-42). In short, the thesis and antithesis of the third antinomy are attempting to offer answers to the question of what explains the fact that in our experience an event (or object or state of affairs—what Kant calls a "happening") is always caused by another.

The notion of an "ultimate explanation" is ambiguous, of course. The proponent of the thesis argument, the rationalist, thinks an ultimate explanation is a *complete* one; while the proponent of the antithesis argument, the empiricist, thinks it is a *unified* one. As we will see, Kant thinks an ultimate explanation is a complete *and* unified one and that the idea of the first is necessary for the second. But first, we should see that the concepts of completeness and unity do indeed represent the central concerns of the third antinomy.

We can begin by considering the thesis:

> Causality in accordance with laws of nature is not the only one from which all the appearances of the world can be derived. It is also necessary to assume another causality through freedom in order to explain them. (A 444/B 472)

The second sentence supports the claim that the thesis is meant to provide an explanation of something, i.e., appearances of the world. Causality through freedom is introduced to explain an appearance, namely, any given appearance of cause and effect.

The premise that there must be a first beginning and the assumption that "completeness of the series" is necessary in order to "sufficiently determine" a cause *a priori*, is the crux of the argument. In order to sufficiently explain any given instance of cause and effect, I must completely explain it, i.e., provide an explanation that has a *beginning*, the explanation of which does not depend on referring to another cause. Before analyzing whether it is true that a sufficient explanation of any effect depends on this notion of completeness, let us get clear about what the proponent of the argument is attempting to get at with that claim.

There are two different ways we might interpret the claim that a sufficient explanation is only one that is complete, in the sense of having a first beginning. One option is to think that the universe itself, as something that is constituted by a series of cause and effects, needs a first cause to get the ball rolling, normally thought of as God. From this perspective, the argument would be something akin to the cosmological argument, wherein God is caused by himself and is thus an instance of an "absolute causal spontaneity beginning from itself," which is what the proponent of the argument thinks we need to posit (A 446/B 474). On this view, God is the first cause necessary to provide any sufficient explanation of any instance of cause and effect in experience. This interpretation leaves open whether or not any subsequent cause is free or determined, but posits that there needs to be at least one instance of a first cause, since otherwise we would not have a sufficient explanation of any "happening." As such, this way of interpreting the thesis argument could include a variety of scholastic and early modern views.

Another option is to think that the proponent of the argument is not arguing something about the whole universe *per se*, but about any type of cause that is not caused. The cause could be a local one, i.e., a human one. Aristotle nicely represents this position:

> … the stick moves the stone and is moved by the hand, which again is moved by the man: in the man, however, we have reached a movent that is not so in virtue of being moved by something else. *Physics*, Chapter 5, Book VIII

Aristotle goes on to argue that there must be a first mover of such a causal chain, since otherwise there would be no first cause. The idea of *completeness*—in the sense that a series has a first beginning—is the crux of this type of reasoning.

If we interpret the thesis argument in this local way—i.e., that there must be a first mover of a causal chain, we can begin to get an intuitive sense of what the completeness requirement amounts to. Suppose we want to explain why the stone moved. The stick moved it. But saying this is not, on this view, a *sufficient* explanation of why the stone moved. For that we need an explanation of why the stick moved. Saying that hand moved it still does not do it; but appealing to the man does. This is because the man is not being moved by anything else. He moves rather than being moved. Importantly, the explanation of why the stone moved is *in*sufficient without appeal to an "absolute causal spontaneity beginning from itself," i.e., the man.

There is more to be said here since the example simply assumes that we need to appeal to a first cause in order to sufficiently explain any given cause and hence seems to beg the question that reference to a first mover is necessary for a sufficient explanation. We could reject the premise that a sufficient explanation depends on reference to a complete series, and indeed, many commentators have done precisely that when analyzing the thesis argument.[106] Norman Kemp Smith, for example, flatly denies that we need to refer to a complete series to sufficiently explain any given instance of cause and effect: "this argument," he says, "cannot be accepted as valid. Each natural cause is sufficient to account for its effect. That is to say, the causation is sufficient *at each*

[106] This is similar, of course, to objections to the cosmological argument.

stage" (1918, 493). I will argue below that this type of objection misses Kant's point, but for now I want to simply note that it is indeed the idea of completeness that drives the thesis argument. Furthermore, freedom is not the thing being explained; it is being posited as necessary in order for us to provide an ultimate explanation of something else: any particular event.

While the thesis argument is driven by the idea of completeness, the antithesis argument is driven by the idea of unification. The antithesis concludes:

> There is no freedom, but everything in the world happens solely in accordance with laws of nature. A 445/B 473

It is unfortunate that the talk of explanation drops out of this conclusion. To mirror the thesis, the antithesis should state that in order to explain appearances of the world, we must refer necessarily and exclusively to natural causes. The argument proceeds, like the thesis, by an indirect proof. It begins by assuming that causality through freedom exists. This would entail that a causal series would "begin absolutely through spontaneity," i.e., a series would exist the beginning of which is not preceded, and hence not determined by, any constant law. But the "unity of experience" is only possible given the fact that every action presupposes a state that in turn can be causally connected with a preceding event. Thus, it is not the case that causality through freedom exists.

The crux of the argument lies in the premise that causality through freedom would violate the *unity* of experience. The unity here refers to each event being causally connected to another in a way that makes each event an event *of the same* possible experience of the same world. The only way to understand them as such is by understanding them as causally related to other events. One event is necessarily

determined by another, and so on. An "unconnected" causal series, i.e., one that starts by itself and is at any point not causally determined by what comes before it, is not in this sense unified with the rest of the world, or the rest of a possible experience. Such an unconnected series would stand alone and separate from the other components of experience, and hence, not be connected in a way that makes it of the same possible experience.

Kant's worry in the third antinomy is that it seems as though we can either have a unified explanation of any given event, by appeal to the laws of nature, or a complete explanation, by appeal to causality through freedom, but not both, at least on a traditional metaphysical picture. On a traditional picture, "the stone moved because *I* moved it" is a complete explanation, but not a unified one. It is "complete" because it begins with me. It is un-unified, though, precisely because of this completeness: if the causal chain *genuinely* begins with me, then it is causally unconnected with whatever comes before, i.e., not determined by any law of nature, and thus, not a part of a unified causal series. But of course, it *must* be genuinely empirically unified and is: for I can offer some empirical explanation of why I moved it. I was playing a game or I was drunk, etc.

The real contradiction of the third antinomy, then, is that in order to completely explain—in a unified way—any given effect, we must necessarily assume freedom *and* we must necessarily assume a view of natural causation that seems to rule it out. Rather than freedom being a possibility to be explained, as on the standard interpretation, the explanatory-problem interpretation asserts that freedom is a necessary but seemingly impossible explanans. So the necessity of freedom must be made compatible with the

fact that another necessity seems to rule it out, i.e., the necessity of determinism. On the standard interpretation, a necessity—determinism—is in tension with a mere possibility—freedom. But clearly, the third antinomy arises because of two *necessities*: freedom and the causal law. Thus, it is closer to a true paradox: it expresses two claims both of which must be true but cannot, it seems, be true at the same time. Only once we recognize this can we see why Kant addresses—as a subservient and secondary point—how freedom and determinism can be reconciled.

4. The Completeness Requirement and Human Action

The "explanatory-problem" interpretation has a burden the standard interpretation does not have: it must explain how Kant might justify the completeness requirement, not just in general, but to the proponent of the antithesis argument. I noted above that many commentators of the first *Critique* simply deny the claim of the thesis, i.e., that we need a complete explanation (an explanation that appeals to a genuine causal beginning) in order to sufficiently explain any given instance of cause and effect. Kemp Smith, for example, says an appeal to a previous cause is sufficient (1918, 493), implying that it is sufficient even if that cause is not the genuine beginner of a causal chain. Likewise, the empiricist could simply say that causal determinism rules the day, that it rules out any type of "complete" explanation, and that we do not need that anyhow, so freedom is not necessary and that subsequently, there is no real problem, and certainly no contradiction. Hence, in order for this type of objection not to be valid, Kant would need to show that even the empiricist cannot reject completeness. This makes the problem significantly

harder to resolve than is thought on the standard interpretation. But as we will see, it illuminates much of what Kant says in the resolution that is otherwise obscure.

Initially, the claim that freedom is necessary in order to provide an ultimate explanation of any particular event seems implausible. Even if human-caused events require reference to a free action, which is not true even according to the empiricist, surely it is not the case that *all* events require reference to freedom. A great many events are indisputably independent of human action, and as such, can be completely explained without reference to such. Atoms move in the clouds and cause rain. O-rings do not expand properly and cause shuttles to explode. Surely Kant would not want to claim that ultimate explanations of rain and explosions must necessarily refer to a free action. Although some rationalist metaphysicians might claim that they do, and appeal to God's freedom as the explanation, *Kant* does not want to do this.

Below I will argue that even these types of events only have explanatory effect precisely because of the idea of completeness, and that correctly interpreting Kant's resolution to the third antinomy illuminates his theory of explanation for these types of events too. But the point is complicated and requires investigation first into how the completeness requirement operates in regards to human action. This is what Kant does in the text. It is easy to forget, partly because the standard interpretation is so ubiquitous and partly because Kant himself seems to get distracted by the free will problem, that the arguments of the third antinomy themselves never mention *human* freedom. Rather, human freedom is only properly introduced in the resolution. Of course, human action is what we care about when we dispute the possibility of freedom (we are not in a quandary

over whether the table is free). But there is a deeper reason Kant introduces human freedom, I believe, that has to do with why the empiricist must embrace the completeness requirement. It is because human action cannot be described in the way the empiricist wants without relying on the idea of completeness. I will now defend this claim.

We can start by recognizing how, according to Kant, the empiricist wants to describe human action. For the empiricist, an event is only properly understood as determined in a law-like way by an event before it. This includes human action. My action of moving a stone has to be not just temporally connected with what comes before it (i.e., of the same temporal continuum), it has to be *causally* connected to what comes before it. Otherwise, the empiricist claims, the "mark of empirical truth, which distinguishes experience from dreaming, would largely disappear" (A 451/B 479). The causal law allows us to distinguish subjective experiences like dreaming, hallucinating, imagining, etc., from objective reality, which has the mark of "empirical truth." The reason, according to the proponent of the antithesis argument, is that "for alongside such a lawless faculty of freedom, nature could hardly be thought any longer, because the laws of the latter would be ceaselessly modified by the former, and this would render the play of appearances, which in accordance with mere nature would be regular and uniform, confused and disconnected" (ibid.). In other words, we can only think of appearances as appearances of nature because they follow each other in regular and uniform ways. If we introduced something "lawless," which the proponent assumes freedom is, then natural laws would be "ceaselessly modified" by this lawless faculty.

Suppose, for example, that it is a law of nature that objects always fall to earth at a rate of 9.8 meters per second per second. But a faculty of freedom—by definition here, lawless—would mean they might fall at that rate *except in certain cases*, and that furthermore, there is no way to incorporate the exceptions into the law, since the faculty that is interfering is lawless. This would amount to saying that it is not a law that objects

always fall at a rate of 9.8 meters per second per second. Indeed, *nothing* would be a law, since any type of regularity would be threatened. But it is this very regularity that allows us to distinguish between objective reality and subjective experiences such as dreams, hallucinations, or imaginings.

We should come back to the point that the empiricist thinks that freedom is lawless. This is important, since one might defend freedom by simply changing this assumption. For instance, the compatibilist might say that we are free to act in ways *compatible with* the laws of nature—not against or in complete disconnection from the laws of nature. Of course we are not free to act in opposition to or in freedom from natural laws, the empiricist will say. I am free to get up from my chair but not free to walk on water. But that I am not free to walk on water does not make my getting up from my chair any more or less free. Defining freedom this way, however, does not capture what Kant thinks is necessary for genuine moral responsibility. For it is not as if only some domains of experience are covered by natural laws. The unity of experience requires that *everything* within our objective experience is determined in a law-like way.[107] In this sense, my not getting up from the chair is not different, in the relevant way, from not walking on water. My inability to walk on water is completely determined by the laws of nature. Likewise, my choice to get up from my chair is also so determined—i.e., if I choose to get up from my chair, there are natural laws determining that action, and if I choose to stay in my chair, there are natural laws determining that action. If by *freedom* we mean that I am not *physically* constrained in my chair, then I

[107] Events within episodes of dreams, hallucinations, or imaginings are not so determined, but the episodes *themselves* are, i.e., which is to say that we assume that such episodes have empirical explanations.

am free. But according to Kant this is not moral freedom but the "freedom of the turnspit." The rationalist of the third antinomy agrees. If we introduce natural laws, the rationalist will say, we are no longer talking about freedom.

We can now return to the issue of how the empiricist wants to describe human action. Human action, since it is part of our experience, must meet the unification requirement—that is, it must be casually connected, in law-like ways, to other events that *determine* it. When we describe one's actions in this way, Kant tells us, we are describing one's "empirical character." Kant's own example of a "malicious liar" can serve as our guide here:

> one may take a voluntary action, e.g., a malicious lie, through which a person has brought about a certain confusion in society; and one may first investigate its moving causes, through which it arose, judging on that basis how the lie and its consequences could be imputed to the person. With this first intent one goes into the sources of the person's empirical character, seeking them in a bad upbringing, bad company, and also finding them in the wickedness of a natural temper insensitive to shame, partly in carelessness and thoughtlessness … In all this one proceeds as with any investigation in the series of determining causes for a given natural effect. (A 554/B 582)

A person's "empirical character" is determined by such things as natural predispositions, genetics, peer group, society, and upbringing. Kant later implies that such things determine how one will act—i.e., that if we knew all these things, plus the laws of nature, we could predict one's actions with *complete certainty* (A 550/B 578). Empirical character is a notion we rely on in our interactions with others: *he is the type of guy who never lies, she is a woman of her word, he will stab you in the back as soon as he gets the chance*. We are likewise fairly good at predicting others' actions by referencing the type of character they have—i.e., what actions they will take based on

their past actions, what we know of their "natural predispositions," the people they choose to spend their time with, their role in society, and their upbringing.

So the empiricist explains an individual's actions by referring to all the different types of things that determine one's empirical characters, factors which are all in-turn determined. *Why did he lie*? He lied because doing so is in his character: he was never taught right from wrong, he is naturally selfish, he fell in with the wrong crowd. He was never taught right from wrong because his parents neglected him, he fell in the wrong crowd because he wanted to be accepted… because, because, because. This is the way the empiricist wants to explain human action. And it leaves no room for a certain type of freedom. Sure, no one physically put a gun to the liar's head, but he was sure to lie, even if imperfect humans could not have predicted it.

Here, the burden is to show why the empiricist relies on the idea of completeness even when describing an individual's actions as utterly determined. Kant gives no clear argument for this claim in the resolution. I will argue below that he does offer an *unclear* argument for it, but without some motivation it is hard to recognize. At first glance, it seems as though Kant says that the empiricist must rely on the idea of completeness because he wants to *blame* the malicious liar:

> Now even if one believes the action to be determined by these causes, one nonetheless blames the agent, and not on account of his unhappy natural temper, not on account of the circumstances influencing him, not even on account of the life he has led previously; for one presupposes that it can be entirely set aside how that life was constituted, and that the series of conditions that transpired might not have been, but rather that this deed could be regarded as entirely unconditioned in regard to the previous state, as though with that act the agent had started a series of consequences, entirely from himself. (A 555/B 583)

The liar cannot be to blame, the passage implies, if the liar did not at some point freely choose to lie, which means he was "entirely unconditioned" in the previous state. Forcing Kant into the contemporary free will scheme reveals a disadvantage here: it makes Kant look as though he is begging the question against the empiricist. In order to hold each other morally responsible, we must be free. But freedom is required for moral responsibility, so we must prove the existence of freedom before we prove ourselves morally responsible. But we cannot do so by appealing to our desire to hold ourselves morally responsible; that is a vicious circle. The empiricist can simply respond that freedom and moral responsibility rise or fall together and that the attribution of empirical character relies on neither; indeed, it is the very thing that it rules out.

Even though Kant implies that the empiricist must embrace freedom because he wants to blame the malicious liar, I want to suggest here that the real emphasis of Kant's reply should not be that the empiricist wants to *blame* him—although this will come into play. Rather, it is because the empiricist wants to blame *him*. The real concern is not the attribution of moral responsibility; it is that one's actions, even when they are merely empirically described, are only intelligible when we have a notion of a person's identity, which comes only if we are guided by the idea of completeness. As I will explain in more detail below, in order to attribute to one an "empirical character," the empiricist takes advantage of a special notion of personal identity in order to impute a person's actions to *that person*, which only comes by thinking of an individual under the idea of completeness.

The claim that empirical character is parasitic on the idea of completeness is not obviously supported in the resolution. It is better understood on an initial level through an unpublished reflection:

> In the appearance of inner as well as outer sense one can never regard oneself as the identical self, even as far as the sensible character is concerned. Only with regard to morality, which is the pure consciousness of our self independent from determination in space and time, does the same subject of free actions under the same laws always exist in everything where we are conscious of our self, and there the whole of our actions is regarded as a unity, and we cannot believe that because we have improved ourselves that we therefore have another personality and cannot **be punished on account of the previous one**, as almost all people believe. Of course, this cannot happen with humans as judges. Likewise **one refers evil to one's childhood** (Rousseau: the story about the ribbon) or also what we have done when drunk. Yet improvement is an experience that the character in us is not so entirely evil. (*R* 5456, Kant's bold)

There is much to discuss in this interesting passage, some of which has to do with the attribution of blame. First we should notice, though, that Kant's claim in the first sentence concerns identity. We will recall that Kant supports the claim that personal identity cannot be constituted by the "appearances of inner sense" in the Paralogisms, where he argues against Locke's theory of personal identity (A 361). Here, though, Kant makes the additional point that one cannot regard oneself as identical even using the appearances of *outer sense* as the mark of identity, even as far as "sensible character" is concerned. "Sensible character" seems to mean the same thing as empirical character: a person's actions insofar as they are described as causally unified with the rest of the world. The malicious liar lied because of his upbringing, his peer group influence, and his temperament. Here Kant says that *that* character cannot be thought of as of an identical self either through inner sense *or* outer sense. The implication is that one's own personal identity cannot be perceived through "sense" at all. Just as the rationalist is wrong to say

that I can infer my own identity just from the fact that I have consciousness of the same "I" throughout my thoughts (Kant's argument in the third paralogism), the empiricist is wrong to say that I can infer my own identity just from the fact that I perceive a subsisting object—perhaps the body—in outer sense.

The next important point in this passage is what Kant thinks will solve the problem. The problem is that we cannot attribute to ourselves an identity through inner or outer sense. What allows one then, to attribute an identity to oneself? The short and unsurprising answer is that morality allows us to. But we are in a special position to notice why: because only there—in morality—do we regard "the whole of our actions" as "a unity." *Wholeness* and *unity* are words associated, of course, with completeness and unification. Recall that the proponent of the thesis argument introduced freedom because freedom allows us to offer a complete explanation of any given instance of cause and effect. Here, the completeness requirement reappears, and again, Kant implies that the empiricist himself relies on the idea of completeness. But an important point has emerged: the empiricist relies on the completeness requirement not, like the resolution implies, because the empiricist wants to attribute moral responsibility. Rather, it is because the empiricist wants to unify one's present and future actions with one's previous self.

This desire for a unified account of one's actions is slightly different from what I discussed above in regards to the antithesis argument. There, the focus was on the requirement for one's actions to be causally connected with other events in the world. Here, a similar requirement comes into play. The focus, however, is not on what makes

one's actions unified with the rest of the world, but rather what makes a set of actions the actions of one and the same person. The empiricist needs to view the malicious lie as an action of the same person who was raised in a particular way, has a certain temperament, etc, because otherwise we cannot say that the liar's empirical character determined him to lie. In order to view the lie as the action of a person who performed certain actions in the past and will perform certain actions in the future, we need to have a sense of the liar as a person as a whole, i.e., as a person who is identical over time, who is currently the same person who did action X in the past, and will be the same person who performs action Y in the future. The next step of the argument is to recognize that we cannot have an idea of a person as a whole in this sense without being guided by the idea of completeness.

There are two ways the idea of a person as a whole trades on the idea of completeness. The first is similar to one we have already seen: a complete explanation of any causal event requires us to think of it as part of a causal series that has a genuine beginning. In a similar way, for us to think of a person as a whole, the special type of identity required for empirical-character attribution, we must think of that person as the origin of certain actions. We will see that in this sense, we must think of a person under the idea of freedom. The second way in which the idea of a person as a whole trades on the idea of completeness is more holistic: we need to think of an individual not only as the origin of a particular action, but also as the same originator of a whole set of actions performed across time. I will argue that the latter is a special notion of personal identity that Kant refers to as one's "intelligible character."

The first way in which the idea of completeness is necessary for the idea of a person as a whole is that it requires us to think of a person as the genuine beginner of a causal chain. Why? Because without doing so, the "I"—i.e., the agent who acts—loses the identity it needs in order to attribute to it an empirical character. Suppose I tell a malicious lie. The empiricist wants to be able to say it was *I*, the one who was raised in a certain way, associated with a certain peer group, and was educated in such and such a way, since that is the only way we can say that all of those aspects of my character *determined* me to lie. But this "I," in some sense, needs to be a whole apart from the all of the causal chains my behavior is part of, since otherwise there would be no way to identify—or disambiguate—certain actions or characteristics as *mine*, as opposed to just events or characteristics of some longer or shorter causal chain, not associated with any person. Otherwise, all of "my" actions are just part of a disambiguated causal chain. We lose the "I."

Another way of understanding this point is that thinking of ourselves as the genuine beginner of a causal series is what allows us to maintain the identity of the "I" who acts. We can see this by recognizing the picture of the world that the unification requirement commits us to. To be "unified" with the rest of the world, every action must be causally connected to another action, which is in turn, causally connected to another, and so on. The picture we end up with—and this explains why we are tempted to represent cause and effects with a series of arrows—is that one event or "happening" is just a member of an infinite causal "chain." It is a commitment of the empiricist, or one who wants to attribute to one an empirical character, that there is *one* person to attribute

that character to. To say that "I" performed action X, i.e., to say that "I" stole the ribbon or that "I" lied maliciously, is to do precisely that: it is to indicate that an event within the infinite causal chain (to continue with what we will see is a misleading metaphor) is *of a person*. Furthermore, to say that *I* did it is to indicate—in *some sense*—that I am the genuine beginner of that causal chain. I say "in some sense" here because as we have already seen, there is an important sense in which action X is *not* the result of a finite causal chain, i.e., one that has a beginning, but rather, a member of an *in*finite causal chain. As we will see, thinking of one as the beginner of a causal chain or as the originator of an action just amounts to thinking of one under the idea of freedom—a point that I will say more about below.

First, we should notice that it is not enough, in order to have an idea of a person as a whole, to think of a person being the originator of this or that single action. We also need the idea of a person being the originator of a whole set of actions across time—i.e., we need the idea of a person being the same originator of certain actions in the past, the present, and the future—a sense of his or her identity across time as the originator of certain action. This is another way in which the completeness requirement guides us. And this is what Kant means in the above passage when he says that in morality the "whole of actions is regarded as a unity." The "whole of one's actions," he goes on to indicate, spans time: it includes what one has done in the past, what one does in the present, and what one does in the future.

The implication then is that all of these actions, i.e., the *complete* set of one's actions, must be regarded as of the same self if we are to attribute to ourselves or others

even an empirical character.[108] If we are to understand the malicious lie as the action of the one who lied, we must conceive of the action as belonging to a whole of other actions, all related in a way that makes them actions of the same person. This idea of a complete set of my actions is what allows me to say that it was *I* –the person who acts now—who did X in the past, and also I who will do Y in the future.[109] And because all of those actions are thought of as actions of the same person, that same person is to blame for all of them. Even if I am religiously converted after the sins of my youth, perhaps after I steal a ribbon, my current self is to blame. And notice here that the emphasis is not that my current self is to *blame*, but that it is *me* who is to blame. My "personality," as Kant says, has not changed.[110] I am to blame because the action of stealing the ribbon is only intelligible as the sinful, malicious act it is—causally determined by certain aspects of my personality and my upbringing—if understood as a member of a complete set of actions and characteristics all related in a way we think of as unified, i.e., understood as being of the same person.

---

[108] Andrews Reath's analysis leads to a very similar point to the one I make here. If we attempt to "understand a rational process in terms of empirical causal laws," Reath says, "we lose the sense of it as rational." Not only can we not think of a rational process as described by empirical laws, but once we do, we lose the sense that there is an agent behind those actions: "But once we think of the [rational] process in these terms [of empirical causal laws], we lose the sense that there is an *agent* who is *drawing normative connections* …" (Reath 2006b, 283).

[109] This point—that I am to blame even for my actions done in the past and even for actions I have done "when drunk"—has interesting implications for a criticism often launched against Kantian ethics. In his moral philosophy, as Christine Korsgaard points out, Kant tells us that freedom and rationality are necessarily linked, seeming to indicate that we only act freely when we act rationally, i.e., that we cannot be responsible for our evil actions (1996, 159). If my analysis is correct however, it is precisely because we want to attribute responsibility to ourselves and others for our evil actions or our past actions, that we need to think of ourselves under the idea of freedom. I will say more about this in the last section of this chapter.

[110] Kant's strange suggestion that "almost all people believe" that we cannot be punished on the account of our previous actions seems to be a reference to a religious belief that one wipes the slate clean after religious conversion or confession.

Likewise, the malicious liar better be the same person (i.e., have the same personality) as the person who was raised in a certain way. Otherwise, the empiricist could not say that the liar lied *because* he was raised in a certain way. We are only able to attribute to one an empirical character if we are already being guided by one's identity, which only comes by having an idea of a complete set of one's actions, which is the idea of one's past, present, and future actions as a unified whole. Thus, it is not just that the lie must be causally connected to the liar's upbringing, it is that the lie and the upbringing must be thought of as connected in a more robust sense: that of being of *one* person, since the latter sense is required for causal connections to exist in the way that the empiricist is claiming.

One will notice here that I refer to the "*idea* of completeness," the "*idea* of a person as a whole," and the "*idea* of freedom." I use the word deliberately in the Kantian sense. As we saw in chapter 2, Kant uses the word "idea" in several different ways. One way is to refer to the representation of things not possibly given in our spatial-temporal experience (A 320/B 377). The "transcendental ideas of reason" are examples of such and, again as we saw in chapter 2, these are ideas that, though they are not "congruent" to any objects of possible experience, can regulate our experience by providing a normative standard by which to organize our judgments. So for us to think of an individual under the idea of a person as a whole is for us to think of an individual as the type of thing that is 1) the originator of certain actions, and 2) the identical originator of certain actions over time. To think of an individual under the "idea" of a person as a whole is to say that the latter does not exhaustively capture the spatial-temporal properties of him or her at

any given time. Indeed, at *no* time is an individual captured exhaustively by the idea of him or her as a person as a whole. To say that we are thinking of an individual under such an idea is to say that the latter guides and regulates how we think of the former.

As I have noted, the first way that we think of an individual under the idea of a person as a whole is to think of that individual being the originator of certain actions. This allows us to think of an entity as an *individual* who acts.[111] This is to think of an individual under the idea of freedom, specifically to think of him or her under the idea of what Kant calls "cosmological freedom," which is the "faculty of beginning a state **from itself**" (A 533/B 561). As Kant goes on to say, "the causality" of this type of freedom "does not in turn stand under another cause determining it in time in accordance with the law of nature" (ibid.). And, just as I have noted, he immediately clarifies that freedom in this sense is an *idea*:

> But since in such a way no absolute totality of conditions in causal relations is forthcoming, reason creates the idea of spontaneity, which could start to act from itself, without needing to be preceded by any other cause that in turn determines it to action according to the law of causal connection. (A 533/B 561)

Calling freedom an idea entails it is not an objective concept, in the Kantian sense. It is not as though there is something that exists that begins a state from itself and we form a concept of that thing. Rather, the idea of freedom guides our actions.

As we have seen, the standard interpretation takes Kant's resolution to the third antinomy to resolve how freedom can be compatible with the fact of determinism. I have shown that the problem Kant is really resolving is the fact that we require explanations that necessarily presuppose both freedom and determination. I have argued that Kant

[111] It is the synchronic identity of the agent.

wants to prove that the idea of completeness, and hence the idea of freedom, is necessary even for a unified explanation, and that this is particularly true in regards to the attribution of empirical character, i.e., that we need to be guided by the idea of a person as a whole in order to even attribute to one an empirical character. Kant's discussion of freedom in the resolution to the third antinomy can now be seen in its proper significance. He does not merely show that freedom is possible given the fact of determinism, as the standard interpretation assumes. Rather, he shows why freedom is necessary and that the only way to make that necessity compatible with the necessity of determinism is to show that freedom is an *idea*.

The force of this last point is made stronger if we view it in light of a possible objection. *Even the empiricist*, the argument goes, must think of an individual under the idea of freedom. One might object here, though, that we are not interested in how we *must think* of an individual, we're interested in knowing *what is true* of the individual. Is the individual genuinely free or not, regardless of how we must think of him or her? And herein lies the rub: for if we are asking this question we have not yet understood the force of Kant's resolution. For it is only under a traditional metaphysical picture that this question arises. And again, this shows a weakness of the standard interpretation. For to ask whether the individual is "really" free assumes—as a default position—that we have a concept of an individual person—as causally determined—without already being guided by the idea of a person as a whole. But what Kant is in a position to say (even though he does not say it so clearly in the resolution) is that once the empiricist makes

attributions of a causal determinism, he is *already* and necessarily being guided by the idea of freedom.

One might object though, that freedom still amounts to nothing more than an *idea*. Even if the empiricist himself needs to think of an individual under the idea of freedom, surely this is not the robust claim we wanted, namely, that one can be causally determined *and* free at the same time. Kant's answer to this objection gets to the heart of his resolution: for he will respond that recognizing that freedom is an idea is *the only way* in which we can be causally determined and free at the same time, and recall, both of these things are necessary for the type of explanation reason demands of us. Indeed, one way to phrase the significance of Kant's point in the resolution is to say that he aims to show that freedom is an idea as opposed to a concept. If freedom were a concept of an object, to verify its existence we would have to find a concrete instance of a person acting freely, i.e., of a person *actually* beginning a causal series from him or herself. But as we have seen, the causal law rules out the possibility that an empirical object can freely begin a causal series. But it does not rule out thinking of freedom insofar as we think of individuals as persons who are responsible for their actions. We must not think of freedom as a concept that will be instantiated by a concrete instance of freedom in the empirical world. Instead, to say that we are free just means that we necessarily think of a person as the originator of certain actions. The important point that we should not lose sight of is that we cannot even think of ourselves as "determined" otherwise. As Kant says, "if we would give in to the deception of transcendental realism [i.e., traditional

metaphysics like empiricism and rationalism], then neither nature nor freedom would be left" (A 543/B 571).

I noted that the second way that we think of an individual under the idea of a person as a whole is to think of that individual as the same originator of a certain set of actions spanning time—i.e., it is to think of him or her as an originator who is identical across time. Thinking of an individual as the originator of certain actions allows us to think of that agent as an "I." Thinking of that person as both the originator of certain actions and as identical over time allows us to understand particular actions as having roots in a person's "intelligible character." Intelligible character differs from empirical character insofar as it is based on what we consider to be a person's free actions. In this sense, it can be in one's empirical *and* one's intelligible character to lie; in the first, we view the lie as determined, in the second, as a free choice for which the liar is morally responsible.[112]

This way of interpreting Kant's notion of intelligible character aligns nicely with the text. Kant begins his discussion of intelligible character by saying that he "call[s] intelligible that in an object of sense which is not itself appearance" (A 538/B 566). Now, the human being, of course, is partly an "object of sense" insofar as he or she exists

[112] We might be tempted—following Allison's characterization of Kant's view as a "two-aspect" view—to characterize these as different aspects of the same action—and to a certain extent, this is accurate . What I have shown, though, is that they are not just two possible views we can take of action X. Rather, viewing the liar as determined to lie is only possible because we view the liar as a person as a whole, and thinking of him as a person as a whole necessitates that we think of him as free. This is precisely what Kant means in the above passage when he says that only in morality can one "regard oneself as the identical self." In regards to the malicious liar, we can only attribute the lie to *him* when we think of him under the idea of freedom. Of course, this is also what allows us to *blame* him for the lie. And we can do so without any worry that he was indeed determined to lie.

as a determinate object in space and time. But, along with thinking of a human being as an object in space and time, the above argument showed that we, the empiricist included, must think of a human being in a way that goes beyond what can be captured by any spatiotemporal object, namely under the guiding idea of that human being as a person as a whole who is the originator of certain actions. We should be clear about why thinking of a human being this way is to think of him or her partly as "that in an object of sense which is not itself appearance." This reiterates what it means to say that we think of an individual under the *idea* of freedom. To think of an individual as the originator of certain actions, recall, is not to imply that we will find an actual beginning of a causal chain—in this sense, a free action is not an "appearance." This is just what Kant means by calling it "intelligible." It is something we necessarily attribute to a human being but that is not captured by any sensible features of a human being.

After noting that "intelligible" refers to that in an object of sense that is not itself an appearance, Kant notes that we will "accordingly form an empirical and at the same time an intellectual concept of [this subject's] causality, both of which apply to one and the same effect" (A 538/B 566). He seems to mean here that when we say the liar lied we can mean both that the liar was determined to lie by his empirical character and that he lied freely, which is to say that the lie was rooted in his intelligible character. Thinking of the liar in the first way, Kant goes on to say, is to identify his character "in appearance," while thinking of him the second way is to identify his "character as a thing in itself" (A 539/B 567).

I should clarify precisely what Kant means by implying that each of us has a character as a thing in itself, since this claim is easily misunderstood. We will recall that when discussing the concept of "noumenon," Kant identifies a positive and a negative sense (B 307). To think of noumenon in the negative sense is to think of an object insofar as it is not sensible. For us, it is to think of the object in merely conceptual terms, abstracted from all spatiotemporal content. To think of noumenon in the positive sense, on the other hand, is to cognize an object through concepts alone. Kant argues that we can think of something noumenally in the negative sense but not the positive sense. Here, a similar distinction applies. We cannot, of course, cognize a thing in itself; this claim is at the heart of Kant's transcendental idealism. With his discussion of intelligible character, however, Kant implies that we *can* think of a thing in itself in a negative sense, i.e., insofar as an object of sense is not exhaustively captured by sensibility and hence insofar as we necessarily rely on what is intelligible to guide our thoughts about that object of sense. Part of Kant's argument in the Resolution is that this is precisely how we think of ourselves and others, as rational agents.

I should sum up the argument so far. To attribute to one an empirical character, that is, a rule of how one will be determined to act based on certain other facts about them, one must view a set of actions or characteristics as unified, i.e., as of the same person. But we do not get this sense of personal identity—of actions or characteristics being of the same person—without the idea of a person as a whole, which is tantamount to being guided by the completeness requirement in terms of human action. This idea of wholeness or completeness is what allows us to identify one person that did X in the past,

does Y now, and will do X in the future, and what allows us to think of actions X, Y, and Z as actions of the same "I." The "I" that acts, even when thought of as determined by empirical causes, is not a genuine "I" without simultaneously thinking of it as a genuine originator of actions, i.e., an "I" who *acted*. This is to think of an individual under the idea of freedom and it is to think of an individual as having an identity that spans time—i.e., a character that captures one's actions insofar as they are thought of as free.

Here we should recall that the arguments of the third antinomy endeavor to explain not just human action, but any causal event, presumably even those we would classify as merely mechanical. So far, we have seen a Kantian argument for how the idea of freedom and the idea of completeness necessarily guide us when thinking about human action. As I mentioned above, it might seem implausible that the same ideas must guide us in regards to mechanistic action. The next section aims to explain how it is plausible that they do.

5. The Idea of Completeness and its Role in Non-Human Causation

In this section, I will argue that it is only because we can act freely that we can decide where and when to stop or start an explanatory chain; a part of an infinite causal chain is only significant because we actively approach the world with particular ends and purposes. As such, those explanations are directly guided by the idea of completeness and indirectly guided by the idea of freedom.

Kant's resolution to the third antinomy shows that completeness or a complete causal chain does not exist in an ontological sense, but rather only in what I will call a pragmatic sense. As I noted above, the completeness requirement amounts to an

explanation of a particular effect having a beginning, the explanation of which does not require reference to a further cause. The stone was moved by the stick, which was moved by the hand, which was moved by the man. "The man moved it" satisfies the completeness requirement, which is met if a causal series or causal chain has a beginning, like the man.

This seems simple enough, but an explanation that refers to the beginning of a causal chain is not all that is implied by a complete explanation. Technically, though this is not emphasized in the thesis argument (nor in Aristotle's remarks), a *complete* causal chain would have not only a beginning, but also an end—it would halt in both directions. This is often taken for granted, but is important to note. The man begins the causal chain, and the stone's movement ends it. But why does the stone's movement end it? Surely, the stone's movement also determines some effect (e.g., movement of the sand), so it is not in virtue of the stone's movement being the *ontological* end of a causal chain that it is part of a complete explanation. We seem to take for granted that a stopping point, insofar as it lies at the "effect" end of the so-called "causal chain," is easy to identify, while the "cause" end is not complete until we have reached a genuine mover. This shows us something important about our explanatory practices: what is being explained, the explanandum, is heuristically thought of as the end—in more than one sense of the word—the "end" of the metaphorical chain, and the aspect of the world it is our purpose to explain.

On the "effect" side of the causal chain, we easily acknowledge that the end depends entirely on our purposes, which is to say that apart from our purposes, there is no

ontological end—no ontological feature of the world that says "explain me!" ("I am the end effect of a causal chain!") Undoubtedly, things often do "call out" for explanation, but only when we are already deeply entrenched in a context and complicated web of purposes. That there is no ontological end on the "effect" side of a causal chain is noteworthy because we often assume the opposite on the other end of the causal chain: that what we identify as a cause does *not* depend on our purposes but rather on some ontological feature of the world—a feature that is the actual beginning of a causal chain. In other words, we have the tendency, along with the proponent of the thesis argument, to think that the "beginning" of the chain on the causal side has a *genuine* beginning, an ontological feature of the world that constitutes the "real" beginning of the causal chain. This is perhaps is understandable, for precisely Aristotle's reasoning; we think something must have "really" gotten the ball rolling, whereas where we "end" (the thing being explained) is by its nature arbitrary in an ontological sense. We tend, in other words, to think that whatever got the ball rolling, unlike the rolling of the ball, *has nothing to do with us*, our purposes, or our "ends."

One of the great advantages of my interpretation of Kant's resolution is that it shows that what counts as a causal beginning of a chain, i.e., the originating cause of an event, has just as much to do with our purposes as does the end of our explanations. This is to say that just as there is no effect that ontologically (i.e., without regard to our purposes) announces that *it* is what needs explaining, there is no cause—no beginner of a causal chain—that ontologically announces that *it* is what ultimately explains any given cause. Identifying what needs explanation and what does the explaining, even on a

causal level, is a matter of human purposes in both cases. And this is because—ontologically speaking—there is never a complete causal chain, i.e., one that has an ontological beginning and an ontological end. And we know this is true because the unity requirement is necessarily satisfied within experience: every event is determined by a previous cause. When we say that an explanation is "complete," which we often do, we must mean something other than any claim that we have identified an actual beginning and end of a causal chain. This is what it means to be guided by the idea of completeness. A complete causal chain is what we might call *pragmatically complete*.

Pragmatic completeness does not depend on any ontological feature of the world. It depends entirely on our purposes in asking for and giving an explanation such as who is asking for the explanation, why they are asking, what their background knowledge is, and a whole host of other deeply contextual factors. "Why did the Challenger blow up?" has so many possible answers to it that it would be impossible to offer a satisfactory answer unless we know the conditions under which the question arises. Undeniably, we tend to think there is a *right* answer—something like Feynman's explanation involving the o-ring. But if we do this it is only because we are already favoring a scientific context. No doubt, the scientific context ought to be privileged in a lot of discourse. But we should not fool ourselves into thinking that this is the "correct" answer because it captures the *ontological* halting point of a causal series. It does not; Kant's whole point is that the unity requirement prohibits such ontological completeness. The o-ring explanation is considered "correct," and hence sufficient or complete in a lot of contexts because it serves our purposes in the context in which we ask for the explanation. Hence,

as we can see, it is the idea of completeness that allows us to engage in the very explanatory practices we engage in everyday.

But of course, to say that an idea of completeness guides our explanatory practices is not the same as saying that our explanations of mechanistic events requires the idea of *freedom*, or depends, in any way, on the explainer being free. Indeed, we now have the resources to show that indeed, such explanations do require the explainer to be free—not in the moral sense we saw above—but in a way Kant identifies as "spontaneous." A thinker needs to be "spontaneous" in order to make purposeful judgments about the world, not just because a thinker requires an identity as an "I," but because one can only approach the world with a purpose in mind when one is *active*. An active thinker must initiate any explanatory chain. Both of these points—that the thinker requires a certain type of identity and that the thinker *acts* when he or she makes objective judgments—are not explicitly argued for in the resolution to the third antinomy but in the Transcendental Deduction (see B 129 ff) (and also importantly emphasized in the Preface to the B edition).

In the Transcendental Deduction, Kant tells us, of course, that the "**I think** must be able to accompany all my representations" (B 131, Kant's bold), which is to say that any judgment must be capable of being thought by an "I," i.e., a thinker that has a singular identity. And this representation, Kant goes on to say, is an "act [*Actus*] of spontaneity … [that] cannot be regarded as belonging to sensibility" (B 132). Relying on the insights of the reflection I quoted above, we are in a special position to notice why this "act"—the act of attaching "I think" on my judgments—must be considered a

spontaneous action, i.e., an action that cannot be captured unless we suppose that a thinker is spontaneous, which is just what makes its events *actions*.

Kant's reasoning for the claim that the thinker is spontaneous parallels my reasoning for the claim that the empiricist must rely on the idea of completeness if the empiricist wants to attribute to an individual an empirical character. Likewise, our ability to identify a thinker empirically, what one does when one says "I think," or "she thinks," depends on us thinking of that individual as more than just a spatiotemporal object. We must think of that individual or ourselves as something above and beyond what can be captured empirically. Just as the empiricist must think of an individual as the beginner of a causal chain in order to attribute to that individual an empirical character, we must think of a thinker as the beginner—the initiator—of his or her thoughts in order to attribute those thoughts to that thinker. And indeed, this is just what I am doing when I attach "I think" to my thoughts: I am recognizing myself as the initiator of those thoughts, and I am doing so in a way that implies that "I" am in some sense separate from the empirical chain of representations I think about.

If we think of this "I" as separate from a unified causal chain, we are thinking of it as spontaneous. Notice again, though, that we should not expect to *find* a thinker outside time in this sense. The claim that we can do so is the rationalist mistake Kant criticizes in the Paralogisms. To describe a thinker as spontaneous is not to attribute to it an ontological property but rather to think of a thinker under an *idea*. Otherwise, we would not be able to *empirically* identify and distinguish thinkers from one another, which of course, we do.

Once we understand the point that a thinker needs to be spontaneous, we are in a position to see that only an active thinker can approach the world with a purpose in mind and be the initiator of an explanatory chain. This is to say that only a world with active, free (spontaneous) thinkers is a world that has complete causal chains. Only active thinkers can offer and ask for explanations. This means that our explanation of even mechanistic causal events presuppose that we are free agents.

This idea of approaching the world actively and with a purpose is emphasized in Kant's preface to the B edition of the first *Critique*. There, Kant discusses metaphysics and whether or not it can take the secure course of a science. In regards to mathematics, Kant points out that:

> [a] new light broke upon the first person who demonstrated the isosceles triangle…. For he found that what he had to do was not trace what he saw in this figure, or even trace its mere concept, and read off, as it were, from the properties of the figure; but rather that he had to *produce* the latter from what *he himself thought* into the object…. (B xi, my emphases)

A moment later, Kant makes a similar point about the physical sciences:

> When Galileo rolled balls of a weight chosen by himself down an inclined plane … a light dawned on all those who study nature. They comprehended that reason has insight only into *what it itself produces* according to *its own design*; that it must *take the lead* with principles for its judgments according to constant laws and compel nature to answer its questions, rather than letting nature *guide its movements* by keeping reason, as it were, in leading strings … (B xiii-xiv, my emphases)

Kant's metaphor in this last passage—being guided by "leading strings" harkens back to the "freedom of the turnspit," of course. Even in our non-moral cognitions, Kant is saying here, we cannot just be passively turned or passively guided. In order to make objective judgments, in mathematics, physics, and indeed philosophy, we must recognize that nature answers *our* questions.

This insight has another higher-order implication for the nature of Kant's critique of metaphysics, for it suggests that the very nature of critique is the activity of recognizing the contributions that our own commitments make to philosophical questions and how those contributions shape our philosophical inquiry. The activity of critique is partly just the activity of recognizing that philosophical problems, questions, and answers do not just arise as significant; rather, they have significance based on our purposes.[113] Mathematics and physics proceeded as secure sciences partly because mathematicians and physicists were able to recognize that the answers they received were shaped by the questions they posed, and that their own inquiry provided the framework of understanding to which their respective sciences led. When Kant accuses philosophers, especially metaphysicians, of being "dogmatic" he means that they do not recognize that their own purposes and commitments are what make their answers significant. Rather, they think philosophical truth floats free from any such human commitments or purposes. Kant's transcendental idealism is just a correction of this false assumption.

6. The Moral Self

We are now in the position to make sense of Kant's ethical claims about the self. The interpretation I offer above illuminates Kant's comments about the self in his ethical theory in four different ways. First, we can see why Kant thinks that a being who must act under the idea of freedom *is* free, for all significant purposes. Second, we can see why Kant says that a being who necessarily thinks of him or herself as a person as a whole is subject to the moral law, and why our moral self is, according to him, our

[113] Much thanks again to Dan Ehrlich for helpful conversations on this point.

"proper" self. Third, we can see why for Kant a person is free even when he or she acts immorally or irrationally. Fourth, we can demolish the traditional criticism that Kant assigns moral praise and blame to an impossible target for moral responsibility, namely a noumenal self.

The first issue is how Kant thinks he is justified in saying that we *are* free. As the criticism goes, Kant's idealism hinges on the claim that we can know nothing about the noumenal realm, which means we cannot legitimately attribute any concepts to objects that are not given in intuition. This dissertation has so far shown that for Kant there is some aspect of ourselves that is not given in intuition. This entails then, that we cannot claim any knowledge of that aspect of ourselves. The problem is that Kant himself seems to claim otherwise when he says, in the second *Critique*, that we can know we are free: "freedom is real, for this idea reveals itself through the moral law" (C2, 5:3). The interpretation of Kant's theory that I have offered in this chapter, supported by the major claims of the other chapters, shows precisely why Kant can claim we are free without any inconsistencies on his part.

The key is to recognize that for Kant, even in his moral theory, freedom is still nothing more than an *idea*. This reminder is important, since it bars the criticism that Kant himself is attributing a concept to the noumenal realm. When Kant says we are free, he does not mean that we can know ourselves to be the type of object that is, transcendentally speaking, free and separate from any causal relationships in the world. He means instead that an entity that necessarily acts under the idea of freedom is no different from an entity that *is* free. As Kant puts it:

> every being that cannot act otherwise than *under the idea of freedom* is just because of that really free in a practical respect, that is, all laws that are inseparably bound up with freedom hold for him just as if his will had been validly pronounced free also in itself and in theoretical philosophy. (G 4:449, Kant's emphasis)

If we knew theoretically that a being as a thing in itself was free, it would be subject to the laws of freedom; there would be no difference between how *that* being is required to act and how a being is required to act when acting under the idea of freedom. It is still true that we are not entitled to claim that we are free insofar as we think of ourselves as noumenal. Kant's point is that that "truth" doesn't amount to anything. Since we necessarily act under the idea of freedom, and since doing so leads to the exact same results, it is a difference without significance. Kant's whole purpose in his negative arguments concerning the self was to show that the threat that we are not free has no significant grip on us. Here, he is able to show that we are free precisely because of that lack of significance.

As I promised in chapter 1, we can now see that Kant's Copernican Revolution has shifted the significance of our questions regarding the self. The significance of the question regarding our own freedom, as it turns out, is not whether or not we are absolutely free in sense apart from any of our commitments; it is, rather, whether or not we can legitimately make good on our moral commitments even in the face of determinism. And the answer to that question is a resounding *yes*, that indeed, we must do so to make sense of determinism. For all the purposes that would ever mean anything to us, then, we are free.

The second ethical issue is why a person who necessarily acts under the idea of freedom is subject to the moral law. So far, I have argued that for Kant we must think of ourselves and others under the idea of completeness. This is to say that we must think of ourselves not only as the origin of our actions but as having an intelligible character that is the same source of our actions over time. In order to do this we must rely on the idea of freedom in the sense that an agent transcends any empirical causal description, and indeed, that any such description already presupposes that we are attributing to that person an intelligible character. We should notice that the argument so far operates with the idea of freedom only in a negative sense; its necessity has been proven by showing that a good explanation of human behavior is insufficient *without* it. Kant's moral philosophy shows the positive requirements of what it means for a person to act under the idea of freedom.

In particular, a person who acts under the idea of freedom is, for Kant, bound to the moral law. The first step in seeing this is to recognize that freedom does not entail that one is acting in a completely lawless way. I have argued that it is only possible to think of one having an empirical character if we presuppose that that individual has an intelligible character, and hence, that it is only possible to think of one as determined if we think of one as a person as a whole, under the idea of freedom. So we necessarily think of a human agent as possessing an intelligible character that explains her actions insofar as they are considered free. Kant's point is that one's intelligible character also follows laws, just of a different type than empirical and deterministic ones. Just because a free self (i.e., the self insofar as it is considered free) need not follow natural laws it is

"not for that reason lawless but must instead be a causality in accordance with immutable laws but of a special kind; for otherwise a free will would be an absurdity" (G 4:446). A will that acts in a completely lawless way is not any freer than one that is completely determined. A free will, then, is one that follows laws but not of the empirical, deterministic kind. Rather, it is a will that is subject to the moral law.

This is why Kant thinks the moral self is the "proper" self (G 4:457 and 4:458). A self, as we have seen, is an entity that cannot help but act and think of oneself under the idea of a person as a whole and hence under the idea of freedom. An entity that necessarily acts under that idea is a self, according to Kant. But notice, this is just to say that an entity who acts under the idea of freedom is a self. So When Kant says that the moral self is one's "proper" self, he really means that this is how one is a self at all. The moral self is the self proper because we must presuppose the idea of it in order to attribute other properties of self-hood, including the property of being determined by one's empirical character or the related property of heteronomy, which is when one acts according to one's inclinations instead of the moral law.

This last point—that a heteronomous will presupposes a free will—illuminates an aspect of Kant's discussion that other interpretations struggle with. Kant's moral theory seems to imply that one who acts heteronomously, i.e., one who acts from one's inclinations instead of the moral law, is not one acting freely.[114] If one is not acting freely, one cannot be morally responsible for one's actions. But this implies that we are

[114] For discussions of this criticism, see Allison, 1990, 227-29; Korsgaard 1996b, 159-160; and Wood 1984a.

only morally responsible for those actions that are moral. To return to our earlier examples, one would not be responsible for stealing a ribbon or telling a malicious lie.

But notice that this interpretation of Kant arises only by treating Kant's moral self as if it were an entity in the world whose actions can be empirically measured. A free agent is not that type of thing—indeed, a free agent is an *idea* (as is a *good will*), which means there is no forthcoming example of one acting freely or not, regardless of whether that action is moral. We can only call an action moral or immoral by already presupposing that we treat the agent as a person as a whole and hence as one whose actions can only be attributed to *him* if we already presuppose that he is free, i.e., that we necessarily think of him under the idea of freedom.

In other words, the commentators who are worried that only a good will is a morally responsible one gets the order of explanation wrong: they think that Kant's argument is that we act and then assess whether that particular action was free or not free. On the contrary, Kant's argument is that we necessarily think of ourselves and others as free and that *this* is what allows us to own those actions as our *own*—i.e., take responsibility for them. This entails that whenever *I act*, I am responsible for my actions. Recall that Kant's main concern with the malicious liar is that we are able to blame *him* for the lie, which we are only able to do when we think of him under the idea of freedom.

This raises the important question, of course, of when *I act*, since the implication is that when I do not act, I am not morally responsible for my actions. This way of putting it, of course, is indelicate: by definition I am not responsible for actions that were not mine. The point is to provide a contrast between behavior that I can own and

behavior that I cannot. The necessary distinction, of course, is between an action, which involves choice and mere physical behavior, which does not. If I trip someone in order to embarrass them, I own my behavior. If I trip someone because I am having a seizure, I do not own my behavior, since it is not in the same sense *my action*. Of course, *I* do it in another sense, but not the sense that is important for moral responsibility. I am morally responsible for the first and not the second. So what is the difference? Kant offers a clear answer: behavior is mine—i.e., behavior is *my action*—when it is based on a maxim, i.e., a rule that I follow when I act.

Now, there is a real question here of *when* behavior is based on a maxim.[115] Must I always be conscious of the rule on which I act? That seems improbable at best and wrong at worst. Must I be able to articulate—correctly—the rule I act on? Kant himself says that we might not know our own maxims and even that we can actively misidentify them in cases of moral self-deception (A 551/B 579 and *Metaphysics of Morals* 6: 441-2). Here, I do not attempt to solve the difficult issue of when I do in fact act on a maxim. My aim is to show that Kant has a principled way of deciding when to attribute to one an action—in which case one is morally responsible for it even if it is immoral—and when to attribute to one behavior that one is not responsible for in the moral sense.

Finally, we are able to address the objection to Kant's theory of the self that is launched most often: that it forces him to attribute moral responsibility to a self that is, by his own lights, unknowable, and that furthermore, a "noumenal self" is not the type of thing that can cause events in the spatiotemporal world. Kitcher offers a canonical

[115] For a good discussion of this issue, see Brewer, 2002.

version of the objection: a "noumenal, unknown self is an impossible target for moral criticism and it is at best unclear how we can know that an unknown self creates the formal characteristics of the phenomenal world" (1984, 113). The interpretation I offer above reveals that Kitcher's criticism is misplaced. Indeed, the only reason we have any target for moral responsibility at all is because we think of a human being under the idea of the self. Kitcher, and other commentators who launch this criticism, gets it exactly wrong: the *only* way we can attribute moral responsibility to an agent is by thinking of it as an entity acting under the idea of a person as a whole. This is just what I proved in section 3 above.

We should notice here the metaphysical picture that Kitcher's criticism implies. The complaint is that we cannot know that an "unknown self creates the formal characteristics of the phenomenal world." This implies that the self, for Kant, insofar as it is noumenal, should be an object to which we can apply concepts—an object that we can locate *in* the world and measure whether or not it is actually producing such and such effect. But this is precisely the picture that Kant argues we must reject; it is *transcendentally unreflective* insofar as it does not acknowledge that certain commitments and interests were the very source of the asserted truths. In other words, we came to the picture of the self insofar as it is noumenal—or the person as a whole—or one's intelligible character—by recognizing that we were committed to this idea as the result of attempting to reconcile different human interests we have, in this case, the interests of understanding humans as agents and understanding the world as a unified causal place. The claim, then, that we have an "intelligible character," or that we are

partly noumenal, needs to always contain this qualification. When we speak of the self insofar as it is noumenal, or our intelligible character, or the moral self, we should always include the qualification that such ways of thinking are just that—ways of thinking. Once we think of such descriptions of the self as having significance without this qualification, we commit the same error Kant criticizes in the paralogisms. Kitcher's criticism, and any like it, stem from the very lack of transcendental reflection that plagues the rationalist and the empiricist. And they indicate that such commentators do not yet fully understand the implications of Kant's negative view of the self for his positive one.

To reiterate, to say that these ideas—the self insofar as it is noumenal, the moral self, or one's intelligible character—are just ways of thinking, does not imply they are fictional, unreal, or delusions. This chapter has shown that these ideas are *necessary*. And not just for the way we understand humans in the world, but for the way we understand the world. But they are ideas nonetheless.

A similar response can be given to Longuenesse, who argues that Kant commits a "paralogism of pure practical reason" with his ethical notion of the self. She attributes to him the following argument:

> (1) A subject that is conscious of its own self-determination is a person (in the rationalist sense: an immaterial substance, conscious of its own numerical identity through time.)
> (2) I, as a moral agent, am conscious of my own self-determination (of giving the law to myself).
> (3) So I, as a moral agent, am a person. (2007, 161)

In other words, she accuses Kant, in his ethical theory, of embracing the very notion of personhood that he criticized in the third paralogism. And indeed, one might think that

this is the view that *I* am attributing to Kant in this dissertation. It is not. Allow me to explain why.

My claim in this dissertation is that Kant thinks we necessarily think of ourselves in much the same way Longuenesse describes here: as an immaterial substance, conscious of its own numerical identity through time. I agree, too, that in some sense it is our consciousness of our capacity to determine ourselves that necessitates us thinking of ourselves as immaterial substances identical through time. In other words, I agree that we must presuppose that we "determine" ourselves, which is to say that we are able to act on laws of freedom and in accordance with the idea of an intelligible character. I reject, however, that this entails Longuenesse's first premise: that this means that I *am* a person, in the rationalist sense. Transcendental reflection entails that the only general principle we are entitled to is that a "person" in this sense is one who must necessarily act according to the idea that one is an immaterial substance identical through time. This does not entail that I *am* such a thing—either in the spatiotemporal realm or in the purely noumenal realm. To think so, is to fail to transcendentally reflect on the sources of our claims.

After the publication of the *Groundwork for the Metaphysics of Morals,* Kant was criticized for making claims about the self that seemed inconsistent with his theoretical philosophy. Such claims include the assertions that the self is free, that we can think of the self as partly noumenal, and that the self can cause events. I have shown here that such claims are not inconsistent with his theoretical philosophy—rather, they represent its full development.

THE SELF AND TRANSCENDENTAL IDEALISM

CONCLUSION

Undoubtedly, the view of Kant's self that I have argued for in this dissertation is idealistic in the sense that it hinges on Kant's *idealism*. The self for Kant is essentially ideal. Human beings necessarily act under the idea that each human being is a free, complete, whole entity, with an identity over time. Thus, a "self" is any entity that necessarily acts under such an idea. But I hope I have made clear that this does not mean that it is unreal or fictional, the way, say, "world peace" is just an idea. Quite the contrary: we can only unify our diverse human commitments by recognizing that a self is not an object or a concept that exhaustively describes an object. For if it was either an object or concept, would we not be able to make good on our moral commitments, nor would we be able to make good on our scientific commitments.

I promised, at the beginning of this dissertation, that my interpretation of Kant would both unify Kant's diverse comments on the self and that it would helpfully contribute to the contemporary debates in self-identity and self-reference. I end by saying a bit more about both. First, I will return to and address the problem of self-consciousness under transcendental idealism—namely, the problem that we are conscious of ourselves only as we appear and not how we exist as things-in-themselves. This is related to a major criticism of Kant's theory launched by Peter Strawson. Next, I will gesture at how Kant's theory as I have interpreted it can help us address contemporary

issues that arise with regards to self-identity and self-reference. This last part will remain a gesture, but will point at further work to be done in the area.

1. Kant on Self-Knowledge and Self-Consciousness

What is one conscious of when one is self-conscious? According to transcendental idealism, one is not conscious of oneself *as* oneself or as a thing in itself. Rather, one is conscious of oneself only as one "appears" to oneself. Kant himself admits that this is a strange implication of his view (B 152-53), and indeed, it is one that my interpretation depends on being true. Here, I shall try to explain why this implication is not the devastating implication that, for instance, Strawson takes it to be.

Strawson is somewhat sympathetic to the Kantian project, especially its skeptical aspects. This is particularly true in regards to Kant's remarks on the self in the Paralogisms. Strawson applauds Kant for his critique of the Cartesian ego but implies that it would have been better without the complications of transcendental idealism (1975, 174). He remarks more generally that Kant's arguments against empiricism and rationalism are "developed within a framework of a set of doctrines which themselves appear to violate his own critical principle. He seeks to draw the bounds of sense from a point outside them, a point which, if they are rightly drawn, cannot exist" (ibid., 12). Strawson implicitly refers to the transcendental aspect of the self here—the self that Kant admits is unknowable. Strawson's frustration is that Kant simultaneously wants a transcendental self to be outside of the phenomenal world *and* for it to carry the burden of being the source of the conditions of experience: "For [Kant] the 'I think' of apperception represents … the tangential point of contact between the field of noumenal and the world

of appearances" (ibid., 173). And furthermore, as Strawson goes on to say, it is because of Kant's transcendental idealism that "*we* appear to *ourselves* otherwise than *we* are in *ourselves*." Strawson is distressed by the "confident use of the first personal pronouns," which he says, is "more than bewildering. It shows the model shaking itself to pieces. After all, it seems a good deal can be known about the noumenal self" (ibid., 174).

Strawson seems worried about two related issues here. First, he is worried about what "I" refers to, particularly when it is meant to capture the "noumenal self." Second, he is worried that on Kant's own terms, we should not be able to assert anything of that "I," but that Kant goes on to assert quite a bit about it. The first worry seems to be motivated by the fact that Strawson wants "I" to refer to *one* thing—a human being (ibid., 168). But on his interpretation of Kant, it looks as though "I" has two ways of referring. In one way, it refers to a spatiotemporal object, a human being. In another, it has what we might call a "transcendental designation."[116] The second, on Strawson's view, supposedly refers to the noumenal self, something we should not be able to refer to at all. Kant then goes on to assert many truths about this subject, the most general of which is that it grounds the conditions of the possibility of experience.

The interpretation I have offered of Kant in this dissertation reveals Strawson's worries to be unfounded on both counts. The first worry is taken care of by revealing that Kant actually agrees with Strawson that "I" refers to *one* thing—the human being who utters it. Indeed, Kant's whole argument shows, contra Strawson's interpretation, how it is that "I" can have one referent. In other words, if Strawson wants a representation of

[116] This is not Strawson's term, but Brook's. But it seems to amount to the same thing (Brook 2001, 11).

"I" to have a unified referent, he too must embrace the strange result of transcendental idealism that we can only know ourselves as we appear. Indeed, once we clearly understand Kant's argument, the seemingly strange result that we do not have a certain type of self-knowledge is not so strange after all, but rather innocuous. In regards to the second worry, we will again see how my interpretation solves it: for Kant's point in saying that we can know nothing of this "transcendental subject" is not that we cannot recognize principles that we must be committed to for the sake of experience but rather that even these do not grant us any substantial knowledge of that "I" as an object, since it is no object at all.

When Kant addresses the puzzle that on his view we can only know the self as it appears and not as it is in itself, he frames it not as a problem but as a *solution* to a problem that emerges on previous views of self-consciousness, particularly those that do not distinguish between the thing in itself and the appearance.[117] In regards to the self, this sort of transcendental realism manifests as a view that does not distinguish between apperception and inner sense. So what problem arises if one treats the appearance of the self as the thing in itself and likewise is unable to distinguish between inner sense and apperception?

The problem is that on such a view the self would be *unfathomable*. This is because such a view would mean that we *appear* to ourselves as a "double" I—the "I" that thinks (reflects) *and* the "I" that is apprehended. But if this is the case, then it would be possible that these two "I"s do not refer to the same thing. In other words, saying that

[117] *Anthropology*, 7:141 fn.

we only have knowledge of the self as it appears, solves a problem of self-identity that arises from saying that we have knowledge of the self as it *is*. This is because as thinkers, we think and we can be thought about (by ourselves). Hence, "I" can refer to the "I" of "I think" and to the "I" that I think about. A transcendental realist view of the self says that however the self appears is how it *is* in itself. But on this view I appear in two ways—as a thinker and as an object of thought. But if this is the case, then I can ask the question of whether or not I stay the same across the thoughts I have of myself. But the very question is nonsense. As Kant says,

> To ask, given the various inner changes within a man's mind (of his memory or of principles adopted by him), when a person is conscious of these changes, whether he can still say that he remains the very same (according to his soul), is an absurd question. For it is only because he represents himself as one and the same subject in the different states that he can still say that he can be conscious of these changes. The human "I" is indeed twofold according to form (manner of representation), but not according to matter (content). Anthropology, 7:141 fn

In other words, the representation "I" does not refer in some uses to me as a human being and in other uses as a pure thinker or noumenal self—the matter or content of the representation does not change. If one insists that the self is as it appears, Kant goes on to say, "and if he pursues this investigation as far as he can, he will have to confess that self-knowledge would lead to an unfathomable depth, to an abyss in the exploration of his nature." In other words, unless we have an "I" that transcends all our thoughts, even about ourselves, we would never be able to know ourselves at all.

Indeed, Kant's point is that in order to think of myself as a human being at all in a fathomable way, the representation "I" must transcend the phenomenal world, since that "I" is what unites and synthesizes all my thoughts, including those about myself. Thus,

according to Kant, there are not two separate contents for the use of "I." He suggests, rather, that the representation "I" is "twofold according to form." What does this mean? It means, I propose, not that an utterance of "I" has two possible referents—one a human being and the other a transcendental subject, depending on what the utterance is—but that we are *conscious* of ourselves in two types of ways.[118] The first is as an individual, with a particular history, particular memories, and a particular body. This is self-consciousness in the way we normally mean it and is what Kant calls "empirical apperception." The second is as a "pure I" and is what Kant refers to as transcendental apperception. This type of self-consciousness does not yield any knowledge of myself—only the first type does that. But it is what allows me to synthesize and unify all of my thoughts, including those about myself (as I am presented to myself).

Importantly, on my interpretation, the two types of consciousnesses are, for all significant purposes, indelibly connected. This is obvious in one direction: empirical self-consciousness presupposes pure apperception. But pure apperception does not have any significance without empirical self-consciousness. This was the whole thrust of Kant's critique of rational psychology. Empirical apperception and pure apperception are two necessary sides of the same coin. Kant's critique of the empirical view of the self is that it cannot capture the self it is committed to without pure apperception. Kant's critique of the rational psychologists is that they claim that pure apperception—the "I

[118] Kant implies this in the same footnote I refer to previously (*Anthropology*, 7:141), when he says "And why does [thinking there is a distinction of the appearance and the thing in itself] not present a double I , but nevertheless a doubled consciousness of this I, first that of mere thinking but then also that of inner perception (rational and empirical); that is, discursive and intuitive apperception, of which the first belongs to logic and other to anthropology (as physiology)? The former is without content (matter of cognition), while the latter is provided with a content by inner sense."

think"—can be significant without an object. This dissertation has shown that both sides are insignificant without the other. What this means in regards to self-identity, and the connection between the mind and body, I will say more about below. First, I will turn to its implications for self-reference.

2. Kant on Self-Reference

When I say "I," to what do I refer? Contemporary literature has offered different types of answers, usually in response to Descartes' answer that "I" refers to an immaterial ego.[119] I have shown that Kant is among those philosophers who think that Descartes is wrong in this regard. So what does the "I" refer to for Kant? The most prevalent view in the literature is implicit in Strawson's complaint that for Kant the "I" has two ways of referring—one to the human being and the other to something like the noumenal self or the transcendental ego. Along with Strawson, Sellars seems to espouse this view when he comments that for Kant "the I which thinks is not, as such identical with the I which runs" (1970, 345).

Interpreting Kant this way indeed makes it appear that he anticipated some of insights about self-reference and self-identification found in relatively recent philosophical literature, most obviously the claims of Sydney Shoemaker that certain uses of "I" involve what he calls "self-reference without identification" and are additionally "immune to error through misidentification." Andrew Brook claims that both of these insights originated in Kant (1994 and 2001). My interpretation shows Brook is wrong. Kant does provide insight about self-reference but not in the way Brook claims.

[119] See especially *Meditations*, IV.

Shoemaker's first insight, the inspiration of which he locates in Wittgenstein, is that there are two ways of referring to ourselves—as a subject or an object (1994, 82). "I think it will rain," is an example of the first, "I am bleeding," an example of the second. Shoemaker recognizes that statements of the first type—though perhaps not incorrigible in every way—are at least "immune to error through misidentification relative to the first-person pronouns" (ibid.). In other words, certain uses of the word "I" cannot in principle misidentify the referent. I might be mistaken that I am the one who bleeds, but I cannot be mistaken that I am the one who thinks it will rain. This idea is closely related to the idea that in such uses of the word "I" we refer to ourselves without identifying ourselves, which is to say that one refers to oneself without routing it through any set of properties that one would ascribe to oneself. Like I said, Brook understands Kant to be the originators of both insights about the self.

For the second—self-reference without identification—Brook relies on Kant's remarks on pure apperception (2001, 13). The best example is from the Paralogisms where Kant remarks that the "I" of "I think" is

> designated only transcendentally through the I that is appended to thoughts, without noting the least property of it, or cognizing or knowing anything at all about it. It signifies only a Something in general (a transcendental subject). (A 355)[120]

Brook takes Kant to mean here that when we use "I" in a way that refers to ourselves as subjects, we do so and can do so without knowing anything about the subject to which it refers—hence, self-reference without identification. In regards to the first—immunity to error through misidentification—Brook seems to rely on Kant's claim that pure

[120] Brook also cites A 382 and A 341/B 399.

apperception is "pure," in the sense it does not refer to an spatiotemporal object at all, but rather serves as the original contrast between subject and object. Therefore, its use cannot misidentify its object, since there is no object (2001, 25).

On both counts, Brook misinterprets Kant, as my interpretation shows. In regards to self-reference without identification, although Kant does in fact say that the "I" signifies a "Something in general," it is clear from Kant's critique of rational psychology that such a "transcendental subject" is nothing that subsists, either identically through time or not. A use of the pure "I" of apperception not only does not *identify* anything, it does not even *refer—a fortiori* to a self. This was Kant's point in the paralogisms: all we can say about such uses of the word "I" is that we must think of it as representing a certain type of thing, but that such a conceptual necessity in no way entails that it *does*. But this point is not even the central blow to Brook's interpretation. For my interpretation makes clear that Kant himself rejects the idea that the use of "I" in an utterance or thought like "I think it will rain," refers *only* to a transcendental subject. In other words, Kant would reject the very distinction that referring to ourselves as thinking subjects is to refer to nothing but the self "insofar as it is noumenal," which seems to be Brook's central assumption.

Recall that even the modified conclusions of the rational psychologist that Kant embraces—that I must *think of myself* as a simple substance identical over time—are only the result of abstractions from how I exist spatiotemporally. The fourth paralogism, as well as the Refutation of Idealism, showed that Kant's view is that we cannot refer to a self at all without also being able to refer to external objects. But doing so is to locate

ourselves in that realm of objects—so already, even when we use the word "I" to refer to ourselves as a subject—we are subjects *in the world*—not *just* transcendental subjects or noumenal selves. Now, recall that Kant does say that the distinction between apperception and inner sense provides us two different ways of being conscious of ourselves: one that serves to identify the particular individual I am in the world and the other as a pure "I." Brook might be picking up on this insight. But even that insight does not entail that I refer *only* to a transcendental subject when I utter or think thoughts that refer to myself as the thinking subject (e.g., *I think it will rain*). Like I said above, pure apperception and empirical apperception are two sides of the same coin. This implies that for Kant, when I utter "I," in any sense, I refer to myself, as a human being.[121]

This same reasoning applies to Brook's contention that Kant is the origin of immunity to error through misidentification. On one level, we can sympathize with why Brook thinks this. If so-called transcendental designation refers to a purely transcendental subject and that transcendental subject is not an object at all (which Kant agrees with), then if I use "I" in that sense, I cannot latch on to the wrong object, which is to say that I cannot misidentify it. But notice that this is only because I don't *identify* anything at all. Brook seems to recognize this and notes that he only finds in Kant a modified version of immunity to error: "What we cannot do [in this way of referring to ourselves, according to Kant] is compare it [the subject] to, contrast it with, one object *rather than another*. If so, awareness of self as subject does not distinguish me from or identify me with anything of which I am aware as an object, anything in 'the world'"

[121] In this respect, Kant's view is much closer to the one Strawson wants to find in Kant but does not.

(2001, 25-26). My interpretation shows that we must deflate this claim even more: that because we in principle cannot use the "I" in the sense necessary to generate this immunity, it is a purely theoretical point. Kant's argument in the paralogisms shows that just because I must represent myself as identical over time in no way ensures that I am representing an object that is in fact identical over time.[122] Again, the "I" for Kant refers to the human uttering it. In the next section I will show what this does and does not mean for self-identity.

3. Kant on Self-Identity

So if the "I" for Kant refers to the human being uttering it, what does this entail in regards to Kant's theory of self-identity? The issue of self-identity is particularly vexing because it seems to be crucially tied to difficult metaphysical questions. If a self is something that is identical over time, what is it that remains identical over time? The thought experiments abound: suppose a brain surgeon removes my brain and puts it in another body. Do *I* now exist in another body, or does *my* body now have another brain? In what way to I persist, if at all? We imagine there to be a metaphysical fact of the matter. We also imagine that answering this question is not just important for the sake of attributing moral responsibility but because we care about our own survival. What does Kant's theory—the way I have interpreted it—imply about such questions?

Kant must bite the bullet on the metaphysical issues and embrace the implication that there is no metaphysical fact of the matter about what happens to one's self-identity

[122] Again, Brooks seems to agree with me on this point, making it a bit hard to pinpoint the general disagreement (2001, 26). My point here is that if we are to find an idea similar to immunity to error in Kant, it is so watered-down that it is not worth mentioning. On this, Longuenesse agrees (2007, 155).

when one's brain and body are separated. In this way, Kant's theory of self-identity is similar to that of Derek Parfit, who argues that self-identity is not what is centrally what interests us when we ask such questions. Rather, we care about moral responsibility or survival, which can be addressed satisfactorily without settling the issue of self-identity (1971). In the same way, Kant does not provide us with a notion of self-identity that captures any absolute metaphysical truth that is significant apart from what our interests and purposes are. But he goes even further than Parfit and implies that acting as though the self is something apart from our commitments is what leads to the difficulties of self-identity to begin with.

For Kant, whether we treat a self as identical to another depends solely on our interests and purposes. Now, there is no doubt for Kant that one of these notions of self-identity trumps the others: the sense of identity we must use for the sake of our moral lives (the one I argued for in chapter 5). As I tried to show there, however, it is not just for the sake of our moral commitments that Kant will privilege this sense of identity. It is also for the sake of our general understanding of the world as the unified, causal place we do.

4. The Fathomable Self

I began this dissertation with a quote from Alexander Pope that Kant himself offered in his lectures on anthropology:

> *Human being, you are such a difficult problem in your own eyes*
> *No I am not able to grasp you.*[123]

[123] As quoted by Kant in *Anthropology*, 7:141 fn.

Kant speaks here of how the human self seems unfathomable to us. The self is the source of deeply difficult epistemological and metaphysical questions. I hope I have shown that Kant recognized the difficulty of such questions. I hope I have also shown that contrary to many interpretations of Kant's theory of the self, his view is not only plausible and appealing, but that it exposes why philosophizing about the self is so difficult. Our capacity to reason is continually tempted to overstep its bounds. And it *must* do this; we presuppose and are committed to a particular picture of the world and ourselves within it. Philosophers, however, have the obligation to be conscious of, reveal, and qualify their claims insofar as those claims are the result of those commitments. Doing so, Kant shows us, never weakens the project. Indeed, it strengthens it by confirming that our capacity to reason can satisfy all of our human concerns.

BIBLIOGRAPHY

Allison, Henry E. 1990. *Kant's Theory of Freedom*. New York: Cambridge University Press.

———. 1996. *Idealism and Freedom*. New York: Cambridge University Press.

———. 2004. *Kant's Transcendental Idealism*. Revised and Enlarged Edition. New Haven: Yale University Press.

Ameriks, Karl. 1982. *Kant's Theory of Mind*. New York: Oxford University Press.

Anscombe, G. E. M. 2004. "The First Person." In *Self-Knowledge*, edited by Quassim Cassam, 140–59. New York: Oxford University Press.

Aristotle. 1984. *The Complete Works of Aristotle*. Edited by Jonathan Barnes. Vol. Two. Two vols. Princeton: Princeton University Press.

Arnauld, Antoine, and Pierre Nicole. 1996. *Logic or The Art of Thinking*. Edited by Jill Vance Buroker. New York: Cambridge University Press.

Aune, Bruce. 1979. *Kant's Theory of Morals*. Princeton: Princeton University Press.

Beck, Lewis White. 1960. *A Commentary on Kant's Critique of Pure Reason*. Chicago: University of Chicago Press.

Bennett, Jonathan. 1967. "The Simplicity of the Soul." *The Journal of Philosophy* 64 (20): 648–60.

———. 1974. *Kant's Dialectic*. New York: Cambridge University Press.

Berkeley, George. 2000. *Philosophical Works*. Edited by Michael R. Ayers. North Clarendon, VT: Everyman's Library.

Bermúdez, José Luis. 1994. "The Unity of Apperception in the Critique of Pure Reason." *European Journal of Philosophy* 2:3: 213–40.

Bird, Graham H. 2000. "The Paralogisms and Kant's Account of Psychology." *Kant-Studien* 91: 129–45.

———. 2006. *The Revolutionary Kant: A Commentary on the Critique of Pure Reason*. Peru, Illinois: Open Court Publishing Company.

Brewer, Talbot. 2002. "Maxims and Virtues." *The Philosophical Review* 111 (4): 539–72.

Brook, Andrew. 1994. *Kant and the Mind*. New York: Cambridge University Press.

———. 2001. "Kant, Self-Awareness and Self-Reference." In *Self-Reference and Self-Awareness*, edited by Andrew Brook and Richard C. Devidi, 9–30. Philadelphia: John Benjamins Publishing North America.

Buroker, Jill Vance. 1981. *Space and Incongruence: The Origin of Kant's Idealism*. Dordrecht, Holland: Reidel.

———. 2006. *Kant's Critique of Pure Reason: An Introduction*. New York: Cambridge University Press.

Cassam, Quassim. 1994. *Self-Knowledge*. New York: Oxford University Press.

Castañeda, Hector-Neri. 1990. "The Role of Apperception in Kant's Transcendental Deduction of the Categories." *Noûs* 24: 147–57.

Cohen, Alix A. 2008. "Kant's Answer to the Question 'What Is Man?' And Its Implications for Anthropology." *Studies in History and Philosophy of Science* 39: 506–14.

Collins, Arthur W. 1999. *Possible Experience*. Berkeley and Los Angeles: University of California Press.

De Gaynesford, Maximilian. 2003. "Kant and Strawson on the First Person." In *Strawson and Kant*, edited by Hans-Johann Glock, 155–67. New York: Clarendon Press.

Descartes, Rene. 1998. *Descartes Selected Philosophical Writings*. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. New York: Cambridge University Press.

Dyck, Corey W. 2009. "The Divorce of Reason and Experience: Kant's Paralogisms of Pure Reason in Context." *Journal of the History of Philosophy* 47 (2): 249–75.

Enfield, William 1791. *The History of Philosophy From The Earliest Periods: Drawn up From Brucher's Historia Critica Philosophaie.* London: Tegg and Company.

Engstrom, Stephen. 2006. "Understanding and Sensibility." *Inquiry* 49 (1): 2–25.

Evans, Gareth. 1982. *The Varieties of Reference*. New York: Oxford University Press.

Evans, J. Claude. 1990. "Two-Steps-in-One-Proof: The Structure of the Transcendental Deduction of the Categories." *Journal of the History of Philosophy* 28 (4): 553–70.

Ewing, A. C. 1924. *Kant's Treatment of Causality*. London: Kegan Paul, Trench, Trubner & Co., Ltd.

Gardner, Sebastian. 1999. *Kant and the Critique of Pure Reason*. New York: Routledge.

Ginsborg, Hannah. 2007. "Was Kant a Nonconceptualist?" *Philosophical Studies* 137 (1): 65–77.

Goldman, Avery. 2012. *Kant and the Subject of Critique*. Bloomington: Indiana University Press.

Grier, Michelle. 1993. "Illusion and Fallacy in Kant's First Paralogism." *Kant-Studien* 83: 257–82.

———. 2001. *Kant's Doctrine of Transcendental Illusion*. New York: Cambridge University Press.

Guyer, Paul. 1980. "Kant on Apperception and 'A Priori' Synthesis." *American Philosophical Quarterly* 17 (3): 205–12.

———. 1987. *Kant and the Claims of Knowledge*. New York: Press Syndicate of the University of Cambridge.

———. 1990. "Reason and Reflective Judgment: Kant on the Significance of Systematicity." *Noûs* 24 (1): 17–43.

———. 1992. "The Transcendental Deduction of the Categories." In *The Cambridge Companion to Kant*, edited by Paul Guyer, 123–60. New York: Cambridge University Press.

Hanna, Robert. 2005. "Kant and Nonconceptual Content." *European Journal of Philosophy* 13 (2): 247–90.

Hegel, Georg Wilhelm. 1989. *Hegel's Science of Logic*. Translated by Miller, A. V. Atlantic Highlands, NJ: Humanities Press International, Inc.

Heidegger, Martin. 1990. *Kant and the Problem of Metaphysics*. Translated by Richard Taft. 4th ed. Bloomington and Indianapolis: Indiana University Press.

Hempel, Carl G., and Paul Oppenheim. 1988. "Studies in the Logic of Explanation." In *Theories of Explanation*, edited by Joseph C. Pitt. New York: Oxford University Press.

Henrich, Dieter. 1969. “The Proof-Structure of Kant’s Transcendental Deduction.” *The Review of Metaphysics* 22 (4): 640–59.

———. 1989a. “Kant’s Notion of a Deduction.” In *Kant’s Transcendental Deduction*, edited by Eckart Förster, 29–46. Palo Alto: Stanford University Press.

———. 1989b. “The Identity of the Subject in the Transcendental Deduction.” In *Reading Kant*, edited by Eva Schaper and Wilhelm Vossenkuhl. New York: Basil Blackwell.

———. 1994. “Identity and Objectivity: An Inquiry into Kant’s Transcendental Deduction.” In *The Unity of Reason*, edited by Richard Velkley, translated by Jeffrey Edwards, 123–208.

Herman, Barbara. 1993. *The Practice of Moral Judgment*. Cambridge: Harvard University Press.

Howell, Robert. 1973. “Intuition, Synthesis, and Individuation in the Critique of Pure Reason.” *Noûs* 7 (3): 207–32.

Hughes, R. I. G. 1983. “Kant’s Third Paralogism.” *Kant-Studien* 74: 405–11.

Hume, David. 1974. “An Enquiry Concerning Human Understanding.” In *The Emipricists*. Garden City, New York: Anchor Press/Doubleday.

———. 2009. *A Treatise of Human Nature*. Edited by P. H. Nidditch and L. A. Selby-Bigge. 2nd ed. New York: Oxford University Press.

James, William. 1950. *The Principles of Psychology*. Vol. 1. 2 vols. New York: Dover Publications, Inc.

Kant, Immanuel. 1971. *Prolegomena to Any Future Metaphysics That Will Be Able to Present Itself as a Science*. Translated by Peter G. Lucas. Manchester: Manchester University Press.

———. 1992. *Lectures On Logic*. Translated by Michael J. Young. The Cambridge Edition of the Work of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1993. *Opus Postumum*. Edited by Eckart Förster. Translated by Eckart Förster and Michael Rosen. The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1995. *Kritik Der Praktischen Vernunft*. Köln: Könemann Verlagsgesellschaft.

———. 1996a. “Groundwork of the Metaphysics of Morals.” In *Practical Philosophy*, edited and translated by Mary Gregor, The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1996b. “Religion within the Boundaries of Mere Reason.” In *Religion and Rational Theology*, translated by George di Giovanni, The Cambridge Edition of the Works of Immanuel Kant ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1997. *Lectures on Metaphysics*. Edited by Steve Naragon and Karl Ameriks. Translated by Karl Ameriks and Steve Naragon. The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1998a. *Critique of Pure Reason*. Edited and translated by Paul Guyer and Allen W. Wood. The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 1998b. *Kritik Der Reinen Vernunft*. Hamburg: Felix Meiner Verlag.

———. 1999. “Critique of Practical Reason.” In *Practical Philosophy*, edited and translated by Mary Gregor, The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 2000. *Critique of the Power of Judgment*. Edited by Paul Guyer. Translated by Paul Guyer and Eric Matthews. The Cambridge Editions of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 2002a. “Prologomena to Any Future Metaphysics That Will Be Able to Come Forward as a Science.” In *Theoretical Philosophy after 1781*, translated by Gary Hatfield, The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 2002c. “What Real Progress Has Metaphysics Made in Germany since the Time of Leibniz and Wolff?” In *Theoretical Philosophy after 1781*, edited by Henry Allison and Peter Heath, translated by Peter Heath, The Cambridge Edition of the Works of Immanuel Kant ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 2005. *Notes and Fragments*. Edited by Paul Guyer. Translated by Paul Guyer, Curtis Bowman, Frederick Rauscher. Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

———. 2007a. "Anthropology from a Pragmatic Point of View." In *Anthropology History and Education*, edited by Robert B. Louden and Günter Zöller, translated by Robert Louden, The Cambridge Edition of the Works of Immanuel Kant, ed. Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

Keller, Pierre. 1994. "Personal Identity and Kant's Third Person Perspective." *Idealistic Studies 24* 2 (Spring): 123–46.

———. 1998. *Kant and the Demands of Self-Consciousness*. New York: Cambridge University Press.

Kemp Smith, Norman. 1918. *A Commentary to Kant's "Critique of Pure Reason."* London: Macmillan and Co., Limited.

Kitcher, Patricia. 1982a. "Kant on Self-Identity." *The Philosophical Review* 91 (1): 41–72.

———. 1982b. "Kant's Paralogisms." *The Philosophical Review* 91 (4): 515–47.

———. 1984. "Kant's Real Self." In *Self and Nature in Kant's Philosophy*, edited by Allen Wood. Ithaca: Cornell University Press.

———. 1990. *Kant's Transcendental Psychology*. New York: Oxford University Press.

———. 1999. "Kant on Self-Consciousness." *The Philosophical Review* 108 (3): 345–86.

———. 2011. *Kant's Thinker*. New York: Oxford University Press.

Kleingeld, Pauline. 1998. "Kant on the Unity of Theoretical and Practical Reason." *The Review of Metaphysics* 52 (2): 311–39.

Korsgaard, Christine M. 1996a. *Creating the Kingdom of Ends*. New York: Cambridge University Press.

———. 1996b. *The Sources of Normativity*. New York: Cambridge University Press.

———. 1999. “Self-Constitution in the Ethics of Plato and Kant.” *The Journal of Ethics* 3 (1): 1–29.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Leibniz, G. W. 1981. *New Essays on Human Understanding*. Edited and translated by Peter Remnant and Jonathan Bennett. New York: Cambridge University Press.

———. 1998a. *G. W. Leibniz Philosophical Texts*. Translated by R. S. Woolhouse and Richard Francks. New York: University of Oxford.

———. 1998b. “Principles of Nature and Grace, Based on Reason.” In *Philosophical Texts,* translated by R. S. Woolhouse and Richard Francks. New York: Oxford University Press.

Leibniz, G. W., and Samuel Clarke. 2000. *Correspondence*. New York: Hackett Publishing Company, Inc.

Locke, John. 1979. *An Essay Concerning Human Understanding*. Edited by Peter H. Nidditch. New York: Oxford University Press.

Longuenesse, Béatrice. 1998. *Kant and the Capacity to Judge*. Translated by Charles T. Wolfe. Princeton: Princeton University Press.

———. 2007. “Kant on the Identity of Persons.” *Proceedings of the Aristotelian Society* London.

Marshall, Colin. 2010. “Kant’s Metaphysics of the Self.” *Philosopher’s Imprint* 10 (8).

McCann, Edwin. 1985. “Skepticism and Kant’s B Deduction.” *History of Philosophy Quarterly* 2 (1): 71–89.

Mendelssohn, Moses. 2004. *Phaedon or The Death of Socrates*. Translated by Charles Cullen. Bristol: Thoemmes Continuum.

———. 2007. *Phädon, or On the Immortality of the Soul*. Translated by Patricia Noble. New York: Peter Lang.

Mendus, Susan. 1984. “Kant’s Doctrine of the Self.” *Kant-Studien* 75: 55–64.

Nietzsche, Friedrich. 2001. *The Gay Science*. Edited by Bernard Williams. Translated by Josefine Nauckhoff. New York: Cambridge University Press.

O'Hagan, Emer. 2009. "Moral Self-Knowledge in Kantian Ethics." *Ethical Theory and Moral Practice* 12 (5): 525–37.

Parfit, Derek. 1971. "Personal Identity." *The Philosophical Review* 80 (1): 3–27.

Paton, H. J. 1929. "Self-Identity." *Mind* 38 (151): 312–29.

———. 1961. *Kant's Metaphysic of Experience*. Vol. 1. 2 vols. New York: The Macmillan Company.

———. 1971. *The Categorical Imperative*. Philadelphia: University of Pennsylvania Press.

Perry, John. 2002. *Identity, Personal Identity, and the Self*. Indianapolis: Hackett Publishing Company, Inc.

———. 2008. *Personal Identity*. Second. Oakland: University of California Press.

Pippin, Robert B. 1987. "Kant on the Spontaneity of Mind." *Canadian Journal of Philosophy* 17 (2): 449–76.

Plato. 1992. *Republic*. Translated by G. M. A. Grube, revised by C. D. C. Reeve. Indianapolis: Hackett Publishing Company, Inc.

———. 2000. *Timaeus*. Translated by Donald J. Zeyl. Indianapolis: Hackett Publishing Company, Inc.

———. 2002a. "Meno." In *Five Dialogues*, translated by G. M. A. Grube, revised by John M. Cooper, Second. Indianapolis: Hackett Publishing Company, Inc.

———. 2002b. "Phaedo." In *Five Dialogues*, translated by G. M. A. Grube, revised by John M. Cooper, 2nd ed. Indianapolis: Hackett Publishing Company, Inc.

Powell, C. Thomas. 1988a. "Kant's Fourth Paralogism." *Philosophy and Phenomenological Research* 48 (3): 389–414.

———. 1988b. "Kant's Fourth Paralogism." *Philosophy and Phenomenological Research* 48 (3): 389–414.

Proops, Ian. 2010. "Kant's First Paralogism." *Philosophical Review* 119 (1).

Rawls, John. 2003. *Lectures on the History of Moral Philosophy*. Edited by Barbara Herman. Cambridge: President and Fellows of Harvard College.

Reath, Andrews. 2006a. *Agency and Autonomy in Kant's Moral Theory*. New York: Oxford University Press.

———. 2006b. "Kant's Critical Account of Freedom." In *A Cambridge Companion to Kant*, 275–90. Malden, MA: Blackwell.

Reath, Andrews, and Jens Timmermann, eds. 2010. *Kant's Critique of Practical Reason: A Critical Guide*. New York: Cambridge University Press.

Reich, Klaus. 1992. *The Completeness of Kant's Table of Judgments.* Translated by Jane Kneller and Michael Losonsky. Stanford: Stanford University Press.

Rescher, Nicholas. 2000. *Kant and the Reach of Reason*. New York: Cambridge University Press.

Robinson, Hoke. 1994. "Two Perspectives on Kant's Appearances and Things in Themselves." *Journal of the History of Philosophy* 32 (3): 411–41.

Rosefeldt, Tobias. 2003. "Kant's Self: Real Entity and Logical Identity." In *Strawson and Kant*, edited by Hans-Johann Glock, 141–54. New York: Clarendon Press.

Sassen, Brigitte. 2000. *Kant's Early Critics.* New York: Cambridge University Press.

Sellars, Wilfrid. 1967. "Some Remarks on Kant's Theory of Experience." *The Journal of Philosophy* 64 (20): 633–47.

———. 1969. "Metaphysics and the Concept of a Person." In *The Logical Way of Doing Things*, edited by Karel Lambert. New Haven: Yale University Press.

———. 1970. "... This I or He or It (the Thing) Which Thinks..." Presidential address presented at the Sixty-seventh Annual Eastern Meeting of the American Philosophical Association, Philadelphia, December 28.

Shoemaker, Sydney. 1994. "Self-Reference and Self-Awareness." In *Self-Knowledge*, edited by Quassim Cassam, 80–93. New York: Oxford University Press.

Spinoza, Baruch. 1992. *Ethics: Treatise on The Emendation of the Intellect and Selected Letters*. Edited by Seymour Feldman. Translated by Samuel Shirley. Indianapolis: Hackett Publishing Company, Inc.

Strawson, P. F. 1975. *The Bounds of Sense*. New York: Routledge.

———. 2008. "Imagination and Perception." In *Freedom and Resentment and Other Essays*, 50–72. New York: Routledge.

Stroud, Barry. 1968. “Transcendental Arguments.” *The Journal of Philosophy* 65 (9): 241–56.

Timmerman, Jens. 2010. *Kant’s Groundwork of the Metaphysics of Morals: A Commentary*. New York: Cambridge University Press.

Van Cleve, James. 1981. “Reflections on Kant’s Second Antinomy.” *Synthese* 47 (3): 481–94.

———. 1986. “Kant’s First and Second Paralogisms.” *The Monist* 69 (3): 483–88.

Walker, R. C. S. 1978. *Kant*. London: Routledge and Kegan Paul.

Walsh, W. H. 1975. *Kant’s Criticism of Metaphysics*. Edinburgh: Edinburgh University Press.

Watkins, Eric. 2009. *Kant’s Critique of Pure Reason: Background Source Materials*. New York: Cambridge University Press.

Wilkerson, T. E. 1976. *Kant’s Critique of Pure Reason*. Oxford: Clarendon Press.

Wilson, Margaret D. 1974. “Leibniz and Materialism.” *Canadian Journal of Philosophy* 3 (4): 495–513.

Wittgenstein, Ludwig. 2000. *Philosophical Investigations*. Translated by G. E. M. Anscome. Malden, Massachusetts: Blackwell Publishers, Ltd.

Wood, Allen W. 1975. “Kant’s Dialectic.” *Canadian Journal of Philosophy* 5 (4): 595–614.

———. 1984a. “Kant’s Compatibilism.” In *Self and Nature in Kant’s Philosophy*, 73–101. Ithaca: Cornell University Press.

———. , ed. 1984b. *Self and Nature in Kant’s Philosophy*. Ithaca: Cornell University Press.

———. 1999. *Kant’s Ethical Thought*. New York: Cambridge University Press.

Wright, Larry. 1973. “Rival Explanations.” *Mind Association* 82 (328): 497–515.

———. 1995. “Argument and Deliberation: A Plea for Understanding.” *Journal of Philosophy* 92 (11): 565–85.

———. 2002. “Reasoning and Explaining.” *Argumentation* 16: 33–46.