

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2024

Kant's Invisible Self

Brennan, Haley

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp019880vv382>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

Kant's Invisible Self

HALEY BRENNAN

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
PHILOSOPHY

ADVISERS:
DESMOND HOGAN
ANDREW CHIGNELL

SEPTEMBER 2024

© COPYRIGHT BY HALEY BRENNAN, 2024.
ALL RIGHTS RESERVED.

Abstract

This dissertation explores the role of the metaphysically real self—the self as a thing in itself—in Kant’s theory of the self and self-consciousness. I defend two central theses. First, that Kant’s theory of the self, personal identity, and self-referring mental states relies on the real self. Second, that the role of the real self in Kant’s theory is well-theorized: the explanation of cognitive and metaphysical relations in which we stand to our real selves are compatible with Kant’s overarching philosophical commitments. In defending these theses, I argue that taking seriously the importance of the real self opens up interesting and rich avenues of applicability for Kant’s theory to live questions about the self and identity.

Acknowledgements

I have spent 6 years in my PhD. In that time, I have been supported by many people, communities, and institutions. I cannot adequately express what this support has meant to me over these years, but I will here attempt to make a start.

The foundation of this support is the Princeton Philosophy Department, which has given me access to the rigorous philosophical education, the vibrant intellectual community, and the resources that were crucial for my academic work. I also want to thank the department's amazing administrators, without whom I would have often been lost: Anna Faiola, Brittany Masterson, Daniel Rusnak, Jason Robinson, Jo Kelly, Joann Zuczek, Keisha Craig, and Nancy Groll.

Over the years, I have benefited from presenting and discussing parts of this dissertation in a range of settings. I am very grateful to Béatrice Longuenesse, Clinton Tolley, Colin McLear, Huaping Lu-Adler, Janum Sethi, Katharina Kraus, Thomas Khurana, and audiences at the University of Oklahoma, the Potsdam Centre for Post-Kantian Philosophy, the annual meeting of the Canadian Philosophical Association, the annual meeting of the UK Kant Society, and the Midwest meeting of the North American Kant Society for their comments and questions.

I have been incredibly lucky to have a series of patient, kind, and brilliant advisors in my time as a university student at every level. These people have been instrumental to the progression of my career and my work. What I note here only scratches the surface of the ways I have benefited from their time and support: every year I recognize new ways in which they have helped me.

I am deeply indebted to Des Hogan and Andrew Chignell for their advice, comments, and criticism that have shaped both this dissertation and me as a scholar. I'm especially grateful to Des for challenging my ideas when they needed to be challenged, but making sure I never lost sight of my own convictions. I'm especially grateful to Andrew for giving me the space (and time, natch!) to spin out my wildest views, and for being instrumental in identifying which of these were worthwhile to pursue. I am very lucky to have benefited from the exceptional guidance of both as I navigate my way through the often-daunting enterprise of Kant scholarship.

I have benefited greatly from the wisdom and erudition of Dan Garber in thinking about all areas of philosophy. I am especially grateful to Dan for co-teaching a class with me at Princeton. During that time, I got to witness first-hand how the scope of his knowledge is matched only by his generosity to students, including myself.

Dai Heide first taught me Kant and was beyond patient and encouraging as I stumbled my way through rough attempts to make sense of what was going on in the text. He is responsible for teaching me how to do scholarship and so giving me the tools to write this dissertation: this would not have been possible without his support during my first foray into graduate work.

Lisa Shapiro has been a paradigm of mentorship since I met her in 2016. I am especially grateful to Lisa for including me in the *Extending New Narratives in the History of Philosophy* project, and for trusting me to put together a podcast for the project. It has been one of the most joyful parts of my academic life to date. For this and for unwavering straight talk about life and philosophy both, I am indebted.

Melissa Frankel taught me my first philosophy class, ever, in my undergrad. I liked philosophy, but I didn't really know what to do with it. She was the one who took the time to encourage and support my interest. Without her, I would never have pursued graduate work in philosophy, and I would not be here today.

In addition to this wonderful group of advisors, I have learned from a series of excellent philosophy teachers at Princeton, Simon Fraser, and Carleton. Special thanks are owed to Mark Johnston, who was one of my generals examiners and who gave me comments I have been thinking about ever since; Grace Helton, Liz Harman, Susan Brison, Hans Halvorson, and Ben Morison, who gave me their time and attention while I did my units; Hendrik Lorenz, who has been a supportive and generous DGS; Owen Ware, who sparked my interest in post-Kantian philosophy; and Holly Andersen, Jenn Wang, Tom Donaldson, and Doug Moggach, who have all encouraged me to pursue philosophy at times when I wasn't sure that it was for me.

I have benefited greatly from being surrounded and supported by a wonderful group of graduate colleagues. Thank you especially to Alice Van't Hoff, Arianna Falbo, Ariel Melamedoff, Bridget Brasher, Claudia Dumitru, Claudia Yau, Eleanor Gordon-Smith, Ellie Cohen, Giulia Weißmann, Joy Shim, Maggie Shea, Merrick Anderson, Olivia Branscum, and Thomas Lambert for conversations and support, both philosophically and personally, over the years.

My research has been enhanced by the valuable insights of graduate students and post-docs working on Kant during my time at Princeton: Alexander Englert, Brendan Kolb, Chris Benzenberg, Connor Tannas, Ethan Millar-Virkutis, Ian McKeachie, Lewis Wang, Mario Schaerli, Maya Krishnan, and Maya Kronfeld. I owe special thanks to Alejandro Naranjo Sandoval, who was my grad mentor when I arrived at Princeton, and Yaz Martens, my mentee in turn. Another special thank you, too, to John Ryan and Erik Pervin, with whom I decided to do a reading group on Kant's *Critique of Pure Reason* in the summer during my undergrad. I don't think anything I said then was right, but I will always remember this as the first time I read Kant.

I met Zoe Ashton and Dzintra Ullis in the very first week of my MA, so I've never been in graduate school without them. That's a very good thing, because I am certain I could not have *done* graduate school without them. I am endlessly grateful for their love and support, and to have them as my companions through the world of philosophy.

Daniel Kranzelbinder makes both me and my work better every single day. He is my sounding board when I am stumped, my editor, the verifier of my translations, the first person to hear my crazy, half-baked ideas, and a source of unwavering support. He has read every part of this dissertation, and every part has been improved by his excellent and incisive comments.

Finally, I would not have been able to get through the PhD without the support and love of my friends and family. Thank you to Becca Brennan, Tommy Brennan, Claire Brennan, Karen Martin, James Brennan, Aurora Clare, Ariana Keith, Mary White, Jenna Portelance, Sarah Veska, and Isabel McMurray for reminding me of other sources of joy when the dissertation felt a little too heavy and a little too heady, and for everything that they bring to my life.

Contents

<i>Abstract</i>	3
<i>Acknowledgements</i>	4
<i>Contents</i>	6
<i>Introduction: Revealing Kant's Invisible Self</i>	7
<i>1. Kant on Consciousness of an Invisible Self</i>	17
1.1. <i>What is at Stake: Kant on Self-Consciousness, the Self, and Personal Identity</i>	19
1.2. <i>The Commitments of Real Consciousness</i>	28
1.3. <i>Feeling Towards as a Model for Explaining Real Consciousness</i>	34
1.4. <i>How to be Conscious of Your Real Self</i>	40
1.5. <i>Conclusion</i>	51
<i>2. Kant's Three Kinds of Consciousness</i>	53
2.1. <i>Two Kinds of Consciousness</i>	57
2.2. <i>Three Different Paths for the Two Kinds View</i>	63
2.3. <i>The Need for a Third Kind of Consciousness in the Transcendental Deduction</i>	72
2.4. <i>Further Evidence for Three Kinds of Consciousness</i>	85
2.5. <i>Conclusion</i>	91
<i>3. Are Two Selves Too Many Selves? Puzzles about the Relation Between Selves in Kant</i>	94
3.1. <i>Clarifying the Main Question</i>	97
3.2. <i>Desiderata for an Account of the Identity Relation Between Selves</i>	103
3.3. <i>Evaluating Two Proposed Models for Understanding the Relation</i>	115
3.4. <i>A Positive Proposal: The Actor/Activity Model</i>	128
3.5. <i>Conclusion</i>	139
<i>Conclusion: Self-Cognition, Self-Representation, and Self-Knowledge</i>	142
<i>Bibliography</i>	151

Introduction

Revealing Kant's Invisible Self

1. The Main Argument

The first lines of the passage from which I draw my title are well-known. Frequently quoted, the conclusion to the *Critique of Practical Reason* begins with the following.

Two things fill the mind with ever new and increasing admiration and reverence, the more often and more steadily one reflects on them: the starry heavens above me and the moral law within me.

Given how well-known these first lines are, it is all the more surprising how little attention has been paid to what Kant says next. He writes,

I do not need to search for them and merely conjecture them as though they were veiled in obscurity or in the transcendent region beyond my horizon; I see them before me and connect them immediately with the consciousness of my existence.

The first begins from the place I occupy in the external world of sense [...].

The second begins from my invisible self, my personality [...]. (KpV 5:162)

This invisible self, and the consciousness we have of it, are the focus of my dissertation.

This passage makes three main claims. First, that there is a *metaphysically real self*: a self that is “invisible” because it cannot be intuited in space and time or represented with the concepts employed to make sense of the external world. Second, that we are *conscious* of that self: there is some mental state that has this self as its intentional object. Third, that understanding our *personal identity* (our existence) involves making sense of the relation between my self as it exists in the external world (the empirical self), and the supersensible real self. My dissertation, on Kant's theory of the self, revolves around these three claims.

A *full* account of Kant's theory of the self would require a thoroughgoing understanding of the metaphysically real self, the experienced or empirical self, their relation to one another, and their joint function. In my dissertation, I make progress on giving such a full account by advancing a reading of the least explored of these elements, viz. the role of the metaphysically real self in Kant. I defend two central theses. First, that Kant's theory of the self, personal identity, and self-referring mental states relies on the real self. Second, that the role of the real self in Kant's theory is well-theorized: the explanation of cognitive and metaphysical relations in which we stand to our real selves are compatible with Kant's overarching philosophical commitments. In defending these theses, I argue that taking seriously the importance of the real self opens up interesting and rich avenues of applicability for Kant's theory to live questions about the self and identity.

2. Existing Approaches to Kant's Theory of the Self

My central theses challenge two popular and long-standing strategies for making sense of the real self in Kant. Historically, Kant's theory of the self and self-consciousness has been taken to be inconsistent, and insofar as it refers to the real self, implausible. Jacobi's original criticism—that Kant's account of the mind paradoxically both relies on and rejects the thing-in-itself—seems most barbed when directed at the theory of the self.¹ In a landmark work, Peter Strawson (1966, 249) declared that attempts to identify the empirical and real selves “traversed the limits of intelligibility” and that any project that set out to overcome this difficulty was a doomed venture.

Kant fails to overcome the difficulties concerning identity because they cannot be overcome. There is no refuge but incoherence from the question how the connection is to be made, in the way of identity, between the natural being, the man, with a mental history

¹ Jacobi (1787) challenged the coherence of the claim that the mind is affected by real entities that are both wholly distinct from it, and (by Kant's stated lights) wholly unknowable to it. As he put the paradox: “without that presupposition [of the thing in itself] I could not enter into Kant's system, but with it I could not stay within it.”

of thoughts, perceptions, and feelings and the supersensible being, with no history at all, [here paraphrasing Kant] “in which the representation of time has its original ground.”

Whether tacitly or explicitly, this condemnation of the real self has structured investigations into the Kantian self since. It is so entrenched that, even as readings of Kant which embrace his metaphysical commitments have gained popularity, they have been largely silent on how these commitments manifest in his theory of the self.²

Here is the main issue. Kant is committed to a philosophical theory he calls transcendental idealism. The hallmark of transcendental idealism is that the only objects that we can intuit are those in space and time, and that those objects in space and time are *appearances*, not things as they are in themselves. Historically, it is taken (although frequently challenged in recent literature) that a consequence of this is that we have *no* acquaintance—knowledge, cognition, perception—with things in themselves. But the real self is described as thing in itself: a self that exists outside of space and time. To readers like Jacobi and Strawson, it seems incoherent that we could have a *self*—that there is a part of our existence and identity—that we could not be acquainted with. Call this the problem of the real self.

There have been two main responses to the problem of the real self. One camp, viewing Kant as clearly committed to the existence of the real self but unable to see how it can be reconciled with the rest of his commitments, concludes that Kant is simply inconsistent on this point and so the real self should be ignored or set aside in our best accounts.³ The second camp argues that the

² With notable exceptions: Ameriks (2000, 2012), Watkins (2005), Marshall (2010), Wuerth (2014), and Jaurnig (2021).

³ Aside from Jacobi and Strawson, Kitcher (1984, 2011) is an example of this strategy. Longuenesse also argues in a way that suggests affinity with this camp, e.g. arguing that Kant, in being committed to indispensability of transcendental freedom for morality, falls “prey to a paralogism of practical reason along just the same lines as the paralogism of pure

'real' self is not the metaphysical entity it appears to be. Kant *is* consistent, but only if we deflate Kant's assertions about the real self. On this interpretation, the real self is a thought entity or an object of the mind: a fiction, a heuristic, a perspective, or an ideal.⁴

Corresponding to the conclusions generated by either camp, standard accounts of Kantian consciousness, cognition, and knowledge do not taxonomize any mental activity that has as its referent the real self. Readers tend to treat—often tacitly, and usually without clear or systematic reason—*some* appeals to <consciousness> or 'being conscious' (the German *Bewußtsein* or one of its cognates) as instances of a technical term and part of a complex and rigorous theory, while treating *other* instances as throwaway remarks, stand-ins for any old mental state and therefore easily paraphrased as or run together with, e.g., 'thinkability', 'belief', or 'faith'.⁵

From both directions, more than 200 years after Jacobi's challenge and almost 60 years since Strawson doubled down, the consensus remains that the best account we can give of the self and self-consciousness in Kant is one that excises or deflates the role of the noumenal or real self. I argue that, despite the hurdles in making sense of a metaphysically real self in Kant's theory, this consensus is mistaken.

theoretical reason of which he was accusing rational metaphysicians" (2017, 157), although at other points she argues along lines that are characteristic of the second camp.

⁴ A recent example of this is Kraus (2020), who argues that what I pick out as the real self is the *idea of the soul*: an idea that provides normative guidance for an individual to form their self and identity as a person. As a general and well-known example of this strategy, see Allison (2004).

⁵ This is all the more striking because all of these epistemic states (thought, belief, faith) have their *own* technical meaning and particular structure for Kant. For more on these, see Chignell (2007).

3. The Strategy

Both excising and deflating the real self in Kant's theory come at a very high cost. They dismiss some crucial claims of Kant's about the self and personal identity, make theoretical discussions of consciousness hard if not impossible to square with practical ones, and render his solution to the problem of reconciling the freedom of the will with natural determination extremely difficult to understand (if not simply misguided).

An attractive exegetical consequence of my dissertation is that we do not need to pay this cost. I argue that Kant is both committed to a (metaphysically substantive) real self, *and* that this commitment is consistent with the relevant philosophical commitments of transcendental idealism. I argue for this over the course of three chapters. Each chapter is intended to stand on its own, exploring and answering a set of questions about the real self in Kant in a distinct domain: cognitive, epistemic, and metaphysical. While each offers an independent argument, the arguments in chapters 1 and 2 offer direct support for one another.

All three chapters share the following two interpretive commitments.

(1) <Consciousness> [*Bewußtsein*] is a technical term that picks out a specific epistemic or mental state (unless otherwise specified or there is overriding philosophical reason to treat an instance of it differently).

(2) Kant's theory of the self and self-consciousness is spelled out across both theoretical and practical texts and can only be fully understood when both are taken together.

While both commitments are principles of interpretive charity (that is, they attribute terminological consistency and coherence of thought to Kant), my dissertation shows that we

should also adopt them on philosophical grounds: they yield that most consistent and rich version of Kant's theory.⁶

A note about terminology. Throughout the dissertation, I generally refer to the self that is a thing-in-itself or a noumenon as the *real* or *metaphysically real* self. By this, I mean to track its status as a thing that exists independently of human intuition. And I generally refer to the phenomenal self, or the self that appears, as the *empirical* self, by which I mean to track its status as an experienced, spatio-temporal entity.

4. Overview of Chapters

Here is some more detail about what to expect in each chapter.

Chapters 1 and 2 offer an account of self-consciousness and of the kinds of consciousness that Kant appeals to. A popular account of consciousness in Kant holds that he appeals to two broad kinds. Call these empirical consciousness (awareness of particular representations), and formal consciousness (the form of reflexivity that makes possible this awareness). While the details of this popular account differ across interpretations, they are unified by a shared assumption, namely that these two kinds of consciousness exhaust Kant's account, i.e., that they are the only kinds he appeals to. The implication of this is that many of Kant's claims about consciousness are either aberrations or inexplicable as they do not conform to these categories. In both chapters, I argue against the assumption and against this implication.

⁶ The majority of work on Kant's theory of the self (implicitly) disavows the second commitment, focusing *either* on theoretical *or* on practical claims about the self, with little to be said about how the theory unfolds across texts and domains. An exception to this is Piper (2013), although her interest in developing a contemporary Kantian metaethical account of the self leads her to understand this unfolding in a different way than I develop here (see, e.g., 41).

In chapter 1, “Kant on Consciousness of an Invisible Self,” I argue that there is a gap in a popular interpretation of Kant’s theory of personal identity: by his own lights, this theory must make reference to a metaphysically real self. But if the real self is to be included in the theory of personal identity, we must be able to identify some direct cognitive connection to this self. I argue that Kant offers us an account of exactly this kind of state. Kant appeals to a kind of consciousness that I call *real consciousness*; it has the metaphysically real self as its intentional object. I characterize and explain this consciousness as a non-representational feeling. In defending this view, I show how a commitment to real consciousness is compatible with Kant’s general metaphysical and epistemic commitments and how it clarifies key elements of Kant’s theory of the self: how our moral and theoretical self-awareness are integrated, and how we can begin to shape a comprehensive theory of personal identity in Kant.

In chapter 2, “Kant’s Three Kinds of Consciousness,” I argue that Kant is committed to three conceptually distinct kinds of consciousness. The argument here builds on the work in chapter 1 in the following way. There, I argued that Kant is committed to a consciousness of the real self. Chapter 2 adds to this by arguing that this consciousness is a distinct, third kind: it is not reducible to either formal consciousness or empirical consciousness. This chapter thus provides a different argumentative route to real consciousness based on the taxonomy and concepts of consciousness. This chapter challenges scholarly consensus that Kant describes and makes use of *two* distinct kinds of self-consciousness. This consensus is rarely explicitly stated as such, but it is driving assumption across much work on Kant’s theory of consciousness, the mind, and the self. Call this the Two Kinds View. Against it, I argue that several key arguments in Kant’s discussions of self-consciousness depend on a *third kind* of consciousness: viz. that there is a kind of consciousness

present in these arguments that the Two Kinds View cannot account for. These arguments are all notoriously difficult to understand, and I argue that this difficulty is the result of an attempt to integrate the properties and function of that kind of consciousness into either of two previously identified kinds of consciousness. The result of this integration is a state with incompatible or contradictory properties, thereby rendering Kant's arguments opaque, if not nonsensical. To avoid this undesirable result, we should understand that kind of consciousness as a third, distinct kind. Since the clarity and comprehensibility of several of Kant's arguments hang on there being three kinds of consciousness, I conclude that there are three, not two, kinds of consciousness in Kant: the Three Kinds View.

In Chapter 3, I examine the metaphysical consequences of the view of self-consciousness I argue for in Kant. "Are Two Selves Too Many Selves?" investigates the self that is the object of real consciousness. In particular, I look at how we should understand the relation that obtains between this metaphysically real self and the empirical self that is an object of spatiotemporal experience. Both seem like candidates for the thing that is *me*: my self, or the referent of 'I'. But if I told you that you were two selves, or two people, your reaction would likely be one of confusion. In so stating, I appear to have made a mistake: philosophical and commonsense intuitions both support the view what it is to be *you* is to be exactly one person or self. These intuitions have prompted a long-standing debate about whether Kant holds that there are numerically one or two selves. Here, I argue that our best reading of Kant's metaphysics of the self is one that conceives of the selves as two entities that stand in an intimate relation, and that this reading—far from being confused or implausible—highlights the applicability of Kant's metaphysical framework for tough practical questions about identity. I show that an account of the identity

relation between selves on Kant's view must satisfy three desiderata: it should explain the unity of consciousness, it should accommodate the two distinct set of identity conditions Kant holds for our empirical existence and for our existence as things-in-themselves, and it should explain how moral responsibility can be ascribed to empirical actions. I use these desiderata to narrow down possible explanatory models for the relationship between the empirical and the real self to three options: substance/accident, whole/parts, and actor/activity. I argue that third model fares best in evaluation, and that it offers a way of making sense of what it means for an agent to identify with two selves.

5. Overall Results

The main results of my dissertation are as follows. Kant's theory of the self and personal identity relies on a consciousness that we have of the real self. This consciousness is best understood as a kind of non-sensible feeling that refers to the metaphysically real self (Chapter 1). Kant's account of the mind and of self-referring mental states depends on three distinct kinds of consciousness and self-consciousness (Chapter 2). That self that we are conscious of in this way, the real self, is a distinct entity from the empirical self. They are not numerically identical, but instead stand in a relation of actor to its characteristic activity (Chapter 3). In establishing these conclusions, I also vindicate the two interpretive commitments with which I started, namely that 'consciousness' is reliably used by Kant to pick out a technical term for a state with a specific and unique epistemic state, and that Kant's theory of the self and self-consciousness finds its full expression across both theoretical and practical texts. And I argue that a properly referring consciousness of the real self (and therefore the existence of the real self) is an important pillar in Kant's arguments.

Overall, my dissertation opens up new ways to think about and develop Kant's theory of the self and personal identity in a way that does justice to all of his commitments about the self, both real and empirical. The theory of the self cuts across a range of philosophical questions which, on my interpretation, Kant self-consciously attempts to bring to bear on each other: what it means to be aware of oneself in thought, metaphysical questions about the criterion for synchronic and diachronic individuation, and practical questions about what features, experiences, and values *I* identify with are all interdependent. Identifying these interdependencies helps uncover important elements of his account of the self. By bringing together theoretical and practical claims, I show how Kant's theoretical writings about the self and self-consciousness provide a framework for an account that is filled in with content in his practical writings. This understanding lends insight into some key parts of Kant's philosophy: his theory of cognition, the possible content of representations, the meaning and extension of the kinds of consciousness he appeals to, the nature of freedom, and the relation that obtains between the appearances and things in themselves. In turn, on the interpretation I propose, Kant's account of the self makes available a distinctive contribution to current debates about the self and identity, especially practical and social identity. Kant's distinctive contribution emerges from a close reading of his works given his own perspective and concerns.

Chapter 1

Kant on Consciousness of an Invisible Self

Kant concludes the *Critique of Practical Reason* with some self-reflection. There are two things, he writes, that fill the mind with “ever new and increasing admiration and reverence”: the starry skies above and the moral law within. What is particularly striking about each of these, he goes on, is their immediate connection “with the consciousness of my existence”. The admiration for the former begins from a consciousness of my existence in the external world. The latter begins from a consciousness of “my invisible self, my personality” (KpV 5:162).

This chapter explains the consciousness that we have of our invisible self. ‘Invisible’, for Kant, picks out a class of things that exist beyond the relational order of nature, an order characterized by space and time, the forms of human intuition. This means, first, that an invisible thing in the sense at issue is an object *outside* of space and time: what Kant calls a thing in itself, a thing that exists independently of human intuition. I refer to what has independent existence in this way as what is *metaphysically real*. Second, it means that, since our forms of intuition characterize the order of nature, and since invisible things exist outside this order, we cannot intuit them. In both outer sense (perception and sensible experience) and inner sense (introspection), I can only intuit objects that belong to the order of nature, viz. only objects that appear in space and time.

The consequence of this is that in describing the *self* as *invisible*, Kant is describing a self that is metaphysically real and is not an available referent for standard modes of consciousness, i.e., perception or introspection. What, then, does it mean to be *conscious* of your invisible, real self?

I argue that this consciousness picks out a distinct mental state that has as its intentional object an agent's metaphysically real self. I call this state *real consciousness*. I argue that real consciousness is analogous to a kind of feeling: it is a specific non-representational cognitive relation that is prompted by our awareness of performing free acts. In so doing, I show how the existence of real consciousness is compatible with Kant's overarching metaphysical and epistemic commitments and restrictions, and how this makes room for the real self in Kant's account of personal identity by understanding the consciousness of the real self as a warranted sense of an intrinsic feature of the self.

To get a first grip on this real consciousness and on how to understand its distinctive nature, consider the following analogy. You are walking alone in a forest and suddenly feel a sense of fear, the kind of fear that is caused by proximate danger. You feel your pulse quicken, and the hairs stand up on the back of your neck, but you do not perceive anything that could be responsible for this feeling. You feel fear, then, but you do not perceive the cause of this fear. In this scenario, what is the intentional object or referent of your feeling?

One candidate is your physical reaction itself: the hairs on your neck, the goosebumps on your arm. But, barring a strange phobia, this cannot be a satisfactory account of the referent of your feeling. You are not afraid of those hairs raising, you are afraid of some *thing* that is causing this reaction. In other words, your awareness of your physical reaction is a symptom of the fact that you are feeling fear *towards* something that you do not represent: some fearsome x. Consequently, there are *two* feelings present in this scenario, each with a distinct intentional object. There is a

bodily feeling, directed at your physical reaction. And there is a *feeling-towards*, directed at the cause of your fear.⁷

I argue that the feeling-towards the fearsome, unseen x is analogous to the consciousness we have of the real, invisible self. I spell out the analogy and its application in detail in §3 and §4. The key is the relationship between what you do represent and the object that causes this representation to occur. The upshot of this view is that Kant's invisible self is properly understood as the object of consciousness in the same way that the unidentified fearsome thing is properly understood as the object of the feeling-towards. I show that this understanding of real consciousness as analogous to this feeling demystifies the state without giving up its substantive metaphysical implications: we stand in a direct cognitive relation to our self as it is in itself.

My main claim challenges a popular interpretation of Kant's theory of the self. This popular view denies that we stand in any direct cognitive relation to our metaphysically self. Accordingly, it concludes that the real self cannot be included in Kant's ontology and his best theory of personal identity. In defending my account, I argue against this conclusion. On my view, we can have our metaphysically real self and be aware of it, too.

§1: What is at Stake: Kant on Self-Consciousness, the Self, and Personal Identity

What, for Kant, makes me the individual person that I am? On a popular view, Kant's answers to these questions rest on two central claims.

Claim 1: A Kantian account of personal identity is determined by the nature and content of our self-consciousness.

⁷ The terminology of 'bodily feeling' and 'feeling towards' come from Peter Goldie who works out the distinction between these two states as feelings with distinct referents or intentional objects in detail in (Goldie 2002).

Claim 2: Kant's account of self-consciousness is limited to two kinds: empirical self-consciousness, and formal self-consciousness.

I argue that any view committed to these two claims runs into a problem. The problem is this: Kant appeals to a metaphysically real self and explicitly claims that this self features in what makes us the person that we are. But the metaphysically real self cannot be the referent of either empirical or formal self-consciousness. If the popular reading—any view committed to both *Claim 1* and *Claim 2*—is correct, then the real self is redundant: a 'self' that ultimately plays no role in Kant's theory of the self and the person.⁸ In that case, there is a tension at the heart of Kant's theory of the self. The popular view attempts to resolve this tension by deflating, rejecting, or ignoring Kant's claims about a metaphysically real self. I argue that there is another solution. We can take seriously these claims and their role in Kant's theory of the self and personal identity by rejecting *Claim 2*: there are *three*, and not two, kinds of self-consciousness in Kant's account.

I begin with some detail on the role of self-consciousness in Kant's account of personal identity, which will help to get clear on what is at stake between my view and the view I argue against.

For Kant, what it is to *be* a person is explained, at least in part, by one's ability to be conscious of oneself. This claim is articulated in the major premise of the Paralogism of personality, where Kant offers a definition of personality or personal identity: "What is conscious of the numerical identity of its Self in different times, is to that extent a person" (A361). While—as I discuss below—Kant does not endorse the application of this definition to the referent of a certain kind

⁸ I take it that there are a diverse set of readings and interpretations who all exemplify the popular view (viz., who are jointly committed to both *Claim 1* and *Claim 2*). Amongst them are Henrich (1994) Mariña (2011) Strawson (1966), Rosefeldt (2003), Longuenesse (2017), Allison (2004), Kitcher (2011), and Kraus (2020). I discuss these and how they differ in their commitment to the popular view below.

of consciousness, it is clear that he *does* endorse a version of the definition itself and therefore the tight connection between an explanation of personal identity—the nature of our self, and how we can recognize our self as the same subject over time and across experiences—and an explanation of the nature and content of self-consciousness.⁹ In other words, the question of personal identity becomes a question about the nature and content of self-consciousness: the answer to ‘what makes me the person that I am’, for Kant, will overlap with the possible referents of self-consciousness. This gives us the following claim.

Claim 1: A Kantian account of personal identity is determined by the nature and content of our self-consciousness.

What does Kant have to say about the way that we are conscious of the self? Kant offers a general description of ‘consciousness’ in his logic lectures, where he writes that consciousness “is a representation that another representation is in me.” (JL33). This suggests at least two ways of understanding consciousness. First, consciousness as a representation whose content is another representation. Second, consciousness that that representation is *in me*—an awareness of that other representation as *my* representation.

Kant’s taxonomy of consciousness confirms this. He identifies (e.g., B133, A107) one kind of consciousness that attaches itself to discrete, spatio-temporal representations and so is entirely explained by the features of those representations. Call this **empirical consciousness**. Empirical consciousness is awareness of particular representations.

⁹ See his claim in the *Metaphysics of Morals*: “A person is a subject whose actions can be imputed to them. Moral personality is therefore nothing other than the freedom of a rational being under moral laws (whereas psychological personality is merely the ability to be conscious of one’s identity in different conditions of one’s existence)” (MM 6:223). While this passage is often read as offering two distinct definitions of personal identity, I take it that it is more plausibly read as identifying two joint criteria on being a person, *viz.* having both moral and psychological personality. See his statement at B134 that I would have “as multicoloured, diverse a self as I have representations” unless I was *conscious*.

What about the awareness that a representation is *mine*, or that there is a subject of my representations? The consciousness of any particular complex representation is identified entirely with the features or properties of that representation. If our account of the subject of representations were exhausted by appeal to empirical consciousness, then that subject would change from representation to representation. For this reason, the consciousness of the identity of the subject of representation cannot itself be a representation or have representational content. The consciousness of the identity of the subject is thereby purely *formal*. Accordingly, Kant will often refer to a kind of consciousness that as the *form* of thought: e.g., A398, A685/B713, A346/B404, B411, B427, A321/B378. Call this **formal consciousness**.

Corresponding to these two kinds of consciousness are two kinds of *self*-consciousness. We are aware, via introspection, of our particular mental states. This is empirical *self*-consciousness, or what Kant will refer to as *inner sense*: an awareness of how I appear to myself as an object in space and time. If we abstract out from these particular representations, we are conscious of the necessary condition on having these particular representations, *viz.* that they can be unified in one thinking subject. Since this subject is a condition on having representations, it cannot be identified with any of the representations it makes possible. And since this set of representations is the totality of my representations, there is *no* representation that the subject of representations can be identified with.¹⁰ Therefore, the consciousness of the subject of representations cannot

¹⁰ See B132-6 and A106-9 for the full argument Kant offers for the possibility conditions on thought. See B132-6 and A106-9. In both instances, Kant argues that what I am calling formal consciousness functions as a *transcendental condition*: it makes representations possible, and so must precede them.

have any significant representational content. In formal *self*-consciousness, then, we can be aware only of the structuring role of the thinking subject: the formal conditions of thinking.¹¹

We have in hand two kinds of self-consciousness: empirical self-consciousness, and formal self-consciousness. While Kant does not assert that the formal and empirical kinds exhaust his concepts of consciousness and self-consciousness, there is some basis to think that his overarching metaphysical and epistemic commitments—that is, his commitment to transcendental idealism—limits the possible kinds and referents of self-consciousness to formal and empirical self-consciousness.¹²

There are two reasons in particular to think that transcendental idealism imposes such a limitation. The first is the way that space and time circumscribe the available objects of intuition. The hallmark of transcendental idealism is that the only objects that we can intuit are those in space and time, and that those objects in space and time are *appearances*, not things as they are in themselves (see A26). An agent's particular mental states—thoughts, judgements, feelings, or desires—follow each other as events in time. Therefore, in reflecting or introspecting on these mental states, I intuit them only as they appear (A37-8, A42).

¹¹ This is typically referred to 'transcendental apperception'. I stay away from this terminology for two reasons: first, because I hold that without a good handle on just what content we are trying to explain, this technical term obfuscates rather than clarifies. Correspondingly, I take it that a consequence of my argument in the dissertation is a need to reconsider what technical terminology maps on to what mental function. See §5 and chap. 2 for further discussion.

¹² He does often refer to pair-concepts of consciousness, e.g., *pure* vs. *empirical* consciousness. However, such pairings cannot motivate that there are only two kinds because the terms and concepts that are paired in this way change across texts. The distinction between pure and empirical consciousness is the clearest of these pairings (see A116, B132, B208), but Kant will also often talk about the consciousness of myself as I am affected, contrasted with consciousness of myself as an intelligence, or as I am active (see GM 4:457, B158). Though it is clear that Kant *does* contrast pure and empirical consciousness, the other contrasts he draws make clear that he does not intend this contrast to exhaust his theory. See chap. 2 for further discussion.

I can, Kant writes, abstract from these particular mental states to consider myself only as their joint subject. But, as he argues in the 'Paralogisms' chapter of the *Critique of Pure Reason*, the abstracted or *a priori* representation of myself as a bare subject is not sufficient to support attributing any further, real properties to this subject. This, then, is the second reason for thinking transcendental idealism limits the possible kinds and referents of self-consciousness to formal and empirical. In the Paralogisms, Kant shows that the "rational psychologist"—a philosopher who argues that we can know real properties of the self via introspection and argument—is trading on a fallacious ambiguity between being conscious of a logical subject and a real subject (A346/B404). What matters for present purposes about Kant's arguments in the Paralogisms is that they demonstrate that he denies *de re* awareness of the self as a real entity by virtue of reflection, and inferences from these reflections, on the nature of 'I', the expression of the unified subject of mental states.¹³

In either *a posteriori* introspection (sensible representation) or in *a priori* conceptual analysis and inferential argument (intellectual representation), Kant limits the possible referents of self-consciousness either to the self as it appears, or to the 'self' as the bare or logical subject of thinking. Since these are the referents of the two kinds of consciousness we have canvassed, it should follow that these two kinds are exhaustive of Kant's account. The result is the second claim that motivates the popular view.

Claim 2: Kant's account of self-consciousness is limited to two kinds: empirical self-consciousness, and formal self-consciousness.

¹³ See Ameriks (2000), Longuenesse (2017), and Proops (2021) for three books that devote several chapters to these arguments.

Combine *Claim 1* and *Claim 2* and you get the following thesis: Kant's account of personal identity is determined by the nature and content of empirical and formal self-consciousness. One variant of this view, found in Henrich (1994) and Mariña (2011), argues that the formal unity of the subject is independently sufficient for consciousness of personal identity. Another line, advanced by Strawson (1966), Rosefeldt (2003), and Longuenesse (2017) holds that our awareness of personal identity is based on the identity of the subject actualized in a physical body. Such a view takes the formal conditions of thinking and awareness of our empirical, embodied existence as joint criteria: awareness of personal identity is a result of their relation. Though strategies differ on how exactly empirical and formal self-consciousness work to supply an agent with awareness of their personal identity, they are unified in their commitment to the claim that the only resources for explaining consciousness of *my* identity—its content and its continuity—are either formal or empirical. Any view that shares this commitment is what I call the 'popular view'.¹⁴

Here is a salient problem for this popular view. Kant appeals to an aspect of the human self *as it is in itself* in many instances. Kant states that he "calls intelligible that in an object of sense which is not itself appearance", and that an "intelligible character" would be a causal law that governs the subject as a thing in itself (A538-42/B566-70), later adding that that there is a non-sensible way that the human knows herself (A547/B575). He talks about freedom as the ability of a human

¹⁴ See also Kitcher 2011, Ginsborg 2013 for accounts of Kantian personal identity that share this commitment. Views that argue that we can cognize our self as an object of inner sense could also be sorted into this camp (see, e.g., Chignell 2017), although they could take the conditions on personal identity to be more rigorous than the conditions on self-cognition. See Kraus 2020, Introduction, for an excellent overview of the scholarship on Kant's account of personal identity. While her interest in the specific question of what the object of inner experience is leads her to sort the literature into somewhat different camps (what she calls 'parity' and 'disparity' views), Kraus and I are in agreement that the views outlined cannot properly account for personal identity in Kant's system, although again for quite different reasons. See 2020, chapter 4 for her arguments against existing interpretations.

being as an intelligible (non-sensible) being (MM 6:226). He speaks of a principle that can determine our *existence* “in regard to a certain inner faculty in relation to an intelligible world” (B430-1). In the passage that I began the chapter with (KpV 5:162), Kant calls this part of the human self our *personality*, a term he equates with personal identity (see A345/B403).¹⁵

These passages collectively make clear Kant’s commitment to there being a *self* that is not an appearance: some part of an agent that is outside of space and time. Further, they make apparent that this self is *the ground of free action*. Kant holds that morality—ascribing moral responsibility to actions—is only possible if those actions are grounded in a free action, an action not caused by an event in accordance with a natural causal law (one that states that every cause must have an antecedent cause, or, in other words, that there can be no antecedently undetermined causality). Since such a causal law governs all spatiotemporal events, the freedom required for morality is only possible if it originates outside of space and time. The possibility of morality, then, depends on an existing entity capable of grounding free actions. Altogether, we get a commitment to the existence of a self that is both referred to as part of what makes us the person that we are, and that is identified as playing an indispensable role in Kant’s metaphysics of morality.

References to a metaphysically real aspect of the self tell against a picture on which formal and empirical self-consciousness exhaust personal identity. The popular view cannot preserve the real self by collapsing it into the empirical or the formal self. This real self cannot be the empirical self—the referent of empirical self-consciousness—because the real self it is not an intuitable

¹⁵ See also A554/B582, B429-31, B407, GM 4:452, Pröl. 4:344, KpV 5:97-89, B155, MM 28:878-9, Anth. 7:141 (remark), A492/B520.

appearance (it is a thing in itself). This real self cannot be collapsed into the formal self (the referent of formal self-consciousness) because then it would cease to be “invisible”: the formal structure of our thinking is available to us through introspection and abstraction. And we cannot integrate the real self into the popular view—as is suggested by, e.g., Kraus (2020) and Dyck (2006)—by claiming that it is a thought entity of some kind: a mere fiction, a heuristic of reason, or a normative prescription for the development of the experienced self. For if the real self was a thought entity, it couldn’t ground free action in the way Kant’s moral theory requires it to.

The upshot is that on the popular view the real self is a redundancy within Kant’s theory of personal identity in the sense that Kant may call it a *self*, but it doesn’t actually feature in how we think or know about ourselves as individuals.

Proponents of the popular view recognize these references. They just take it that the exegetical and theoretical cost of declaring the metaphysically real self a redundancy is one that we must pay to be able to offer an interpretation of Kant’s theory of the self consistent with both his transcendental idealism and a plausible theory of the person. Here is the line of thought: if we cannot be aware of the self except as it appears or as it structures our thinking, then the real self cannot be properly included in our best account of personal identity. Given this, we must pay this interpretive cost and ignore, deflate, or reject the role of the metaphysically real self for Kant.

Recognizing this line allows us to formulate the main challenge for the real self, which the remainder of this chapter responds to: if the real self is to be included in the theory of personal identity (and the exegetical and theoretical cost of redundancy thus avoided), we must be able to identify some direct cognitive connection to this self. To avoid further inconsistency, this

cognitive connection must be consistent with transcendental idealism, specifically with the epistemic strictures that it imposes on our cognitive access to things in themselves and the content of our representations.

I argue that Kant offers us an account of exactly this kind of state. There is a *third* kind of self-consciousness in Kant, one that has the real self as its intentional object. I turn now to identify this third kind and gather a set of commitments that an account of it must explain.

§2. The Commitments of Real Consciousness

Kant appeals, in several instances, to a *consciousness* an agent has of her self-in-itself. He never defines this consciousness, but in canvassing passages where it is mentioned and the context of its use, we can put together a set of core commitments that an explanation of this consciousness must account for. In accounting for these commitments, I argue, we can see that Kant has a picture where there are three kinds of self-consciousness: agents stand in a direct cognitive relation to their self in itself in the way required to include this self in an account of the mind and the person.

Here are several passages where this consciousness is appealed to. I bold the relevant parts.

But the very same subject, being on the other side **conscious of himself as a thing in itself**, also views his existence insofar as it does not stand under conditions of time and himself as determinable only through laws that he gives himself by reason; and in this existence of his nothing is, for him, antecedent to the determination of his will, but every action – and in general every determination of his existence changing conformably with inner sense, even the whole sequence of his existence as a sensible being – is to be regarded in **the consciousness of his intelligible existence** as nothing but the consequence and never as the determining ground of his causality as a noumenon. (KpV 5:97-8)

For, that a thing in appearance (belonging to the world of sense) is subject to certain laws from which **as a thing or a being in itself it is independent** contains not the least contradiction; that he must represent and think of himself in this twofold way, however, rests as regards the first on consciousness of himself as an object affected through the senses and as regards the **second on consciousness of himself as an intelligence, that is,**

as independent of sensible impressions in the use of reason (hence as belonging to the world of understanding). (GM 4:457)

Sometimes Kant presents this consciousness of a real self as co-extensive with a consciousness of freedom or spontaneity, i.e., a power to bring forth a state from oneself without an antecedent temporal cause. Here, he equates awareness of the two:

For, the sensible life has, with respect to the **intelligible consciousness of its existence (consciousness of freedom)**, the absolute unity of a phenomenon, which, so far as it contains merely appearances of the disposition that the moral law is concerned with (appearances of the character), must be appraised not in accordance with the natural necessity that belongs to it as appearance but in accordance with the absolute spontaneity of freedom. (KpV 5:99)

Here, he claims that consciousness of spontaneity can supply the recognition of myself as real:

[...] but with the qualification that he can only recognize himself as an object through his representation in experience as he appears to himself, not as he, the observed, is in himself. – If he wished to cognize in the latter way, he would have to rely on a **consciousness of pure spontaneity (the concept of freedom)**, (which is also possible), but it would still not be perception of his inner sense (inner experience) which is based on it. (Anth 7:141n)

This is echoed in the *Critique of Pure Reason*: B157-8 and a corresponding footnote where Kant describes a consciousness “only that I am” and claims this is a consciousness “of my spontaneity”, which gives me “reason to call myself an intelligence”.¹⁶ These claims are coupled with several other points in the text where Kant appeals to a consciousness that seems to be of my real self: B429 refers to an unknowable “separate existence of the object of pure consciousness” and A546/B574 claims that human beings know themselves as “intelligible objects” through “mere apperception”.¹⁷

¹⁶ See too the *Critique of Practical Reason*’s assertion that I am conscious of my invisible self (KpV 5:162), and the claims at KpV 5:6n and B429 that we are conscious of ourselves as the “being itself”.

¹⁷ As I said in n.11, ‘apperception’ is an important technical term in Kant’s account of self-consciousness. This passage offers a good opportunity to motivate my use of ‘formal’ or ‘real’ consciousness in its place for the purposes of *this* chapter. Sometimes, Kant uses ‘apperception’ to pick out what I distinguish as ‘real consciousness’, and sometimes, he uses ‘apperception’ to refer to the state I distinguish as ‘formal consciousness’. Until these distinctions are well-

In every example, Kant appeals to a *consciousness*—the German *Bewußtsein* or one of its cognates—where the content of that consciousness is either the self (Selbst) as a thing in itself, or it is pure spontaneity or freedom. Taking these claims at face value, Kant is describing a kind of awareness that an agent has of herself as metaphysically real: a supersensible subject. This state is what I call **real consciousness**.¹⁸

So what *is* real consciousness? What kind of mental access is Kant claiming that we have to our real self? From the passages, we can identify two commitments.

C1: The content of the state—its intentional object—is the real self.

C2: The content of the state is related to an agent's awareness of freedom.

I pull these apart as two separate commitments even though Kant draws or at least suggests an equivalence between these (that consciousness of freedom is consciousness of the real self, and vice versa) because an explanation of real consciousness should account for this equivalence. The remainder of the section identifies four further explanatory commitments for an account of real consciousness.

The third and fourth commitment both emerge from the considerations surveyed in §1. If we are to properly situate real consciousness in Kant's account, its nature must be understood such that it does not violate Kant's overarching philosophical commitments. In particular, it must be compatible with his epistemic and cognitive strictures. Taking seriously the consciousness we

understood and firmly in place, references to apperception will result in a confusion between the two kinds. I take it that the work I do here results in the ability to better comprehend the distinction in uses of apperception.

¹⁸ I refer to mental 'states' throughout the chapter. This is a terminology choice for clarity and simplicity and does not represent a substantive claim about the ontology of mental things for Kant. There may be good reason to properly term these 'states' as activities or capacities: if those descriptors appeal to you more, you can substitute them in for 'state' with no changes to the substance of my argument.

have of our real self means taking seriously a kind of consciousness which directly *refers* to a supersensible subject. In order for this to be compatible with the strictures imposed by transcendental idealism, the grip that we have on the real self via consciousness must be distinct from either sensible representation (since we cannot intuit something that is not in space or time), or from intellectual representation (since, as the Paralogisms make clear, the path of intellectual representation does not confirm the reality of its representations). If consciousness of the real self cannot come via either sensible or intellectual representation, then it cannot be a representation at all (at least not a direct representation).¹⁹ But the real self is clearly the intentional object of the state of real consciousness. Explaining the compatibility of real consciousness with Kant's epistemic restrictions, then, amounts to solving a puzzle: how a mental state can have an intentional object that it does not directly represent. This is the third commitment.

C3: Real consciousness can be explained as a mental state with an intentional object that the haver of the mental state does not have a direct representation of.

Likewise, both the content and the form of real consciousness must be conceptually distinct from the previously two identified kinds of consciousness (formal and empirical consciousness). Otherwise, Kant's arguments for the nature of either of these two kinds would be in jeopardy: we would either lose the specific intuitive content that characterizes empirical consciousness, or the formal content that characterizes formal consciousness. From this, we get a fourth commitment:

C4: Real consciousness is a third, conceptually distinct kind of consciousness.²⁰

¹⁹ Kant contends that we can have representations of particular intuitions, and representations of concepts that lack any possible corresponding intuitions (like the ideas of reason). He argues, in the Transcendental Dialectic, that we would only be justified in taking these conceptual representations to be referring by way of a rational argument, and that these arguments (at least in the case of the concept of the self) are fallacious. Hence, we need a non-representational route.

²⁰ I defend this commitment in chapter 2.

A fifth commitment concerns the relation between this consciousness and an agent's awareness of the moral law. In his *Critique of Practical Reason*, Kant asserts that rational human beings are immediately conscious of the moral law, and that this immediate consciousness offers a justification of our freedom or spontaneity that we are unable to attain via theoretical reason. Freedom and the moral law stand in a reciprocal relationship where freedom is the "ratio essendi" of the moral law, and the moral law is the "ratio cognoscenti" of our freedom (KpV 5:5). That we are free is the explanatory and metaphysical ground of the applicability of the moral law to human actions, but the moral law has epistemic priority: our knowledge of our freedom comes by way of the moral law. Several of the instances where Kant refers to the consciousness we have of our real self are in the course of discussions of precisely this reciprocal relationship between the moral law and freedom. It is taken, then, that whatever consciousness we have of ourselves as real entities (our real selves) it is closely related to our awareness of the moral law, to the point of regularly being entirely identified with it.²¹

While I agree that there is a close connection between the moral law, human freedom, and our awareness of our real selves, our awareness of the moral law cannot constitute or exhaust our awareness of our real self. There are two reasons for this. First, our awareness of the moral law does not specify *the occasion of the occurrence* of this consciousness.²² As evidenced by my survey

²¹ The strong version of this thesis holds that there is no question to be asked about consciousness in the practical philosophy *except for* a question about consciousness of the moral law. For examples of this strong thesis, see Korsgaard 2009, Kleingeld 2010. A slightly weakened version of this thesis takes it that any investigation into self-consciousness as described in the practical works resolves into a claim about consciousness of the moral law. See, e.g., Kitcher 2021.

²² A consequence of my reading of real consciousness and Kant's theory of the self is that I reject the view, popularized by Karl Ameriks, that Kant originally thought he could offer a proof of theoretical freedom (in the *Groundwork* and the A-Edition of the first *Critique*), before realizing these arguments were insufficient and changing course in the B-Edition and the second *Critique* to a deduction of freedom via the moral law (Ameriks 2000, 208-213). I argue that this view rests on a confusion between two theses that Kant maintains consistently in the Critical period: that we are *conscious* of

of the passages where a consciousness with this content is referred to, this consciousness of myself as thing in itself is not exclusively brought up in connection with the moral law. In fact, it is referred to as a pre-existing state that is clarified when we become aware of the moral law—e.g., in the General Remark at the end of the Paralogisms, where Kant writes that the moral law is an occasion for us to *attribute* determinate detail to the consciousness of our existence (B430-1). For this reason, I contend that consciousness of the real self and consciousness of the moral law should not be identified one-to-one.²³

Second, if we are to successfully offer a complete theory of Kantian personal identity—where completeness means that it can include the real self—the consciousness of our real self should not be restricted to moral cases, even if it finds its full flourishing there. Recall that a problem for Kant seems to be that we can offer an account of identity without appeal to this self, even though Kant takes himself to be justified in calling it a *self*. One reason this problem gets entrenched is because Kant's theoretical and practical arguments are taken to be entirely divorced. However, if we can successfully give a *domain-neutral* account of real consciousness, this gives us room to think that the real self is not restricted to moral cases. Together these reasons introduce what I'll call a *neutrality* commitment.

C5: The explanation of real consciousness is neutral, i.e., it can be properly integrated into both its theoretical and practical contexts without presupposing or excluding the other.

our real self and our freedom, and that we *know* our freedom through the moral law. A full defense of this argument is beyond the scope of this chapter. See Choi 2019 for an argument along similar lines, *viz.* that Kant's claims about consciousness preclude the possibility that he changes his mind about this direction of argument.

²³ Even if you remain committed to their equivalence, you should still accept that appeal to a consciousness of the moral law is not in and of itself explanatory. If you ask, 'what does it mean to be *conscious* of my real self?', and I answer is 'it means you are conscious of the moral law', I have merely glossed consciousness of one thing in terms of another: I have not told you what the cognitive relation in question is. I am here concerned with explaining consciousness of the real self. If that does amount to consciousness of the moral law, then that explanation can be extended.

The sixth and final commitment is that the explanation of real consciousness is compatible with thinking of the self as a distinct, supersensible subject.²⁴ That is, substantiating the consciousness that we have of our real self is a defense of the claim that there is a metaphysically real subject that we are aware of and that we identify with. The final commitment is to the cogency of these claims: an account of real consciousness should justify how an agent can have an awareness of their self as *both* a spatiotemporal object *and* as real, i.e., as something beyond or more than the sum total of their spatiotemporal experiences.

C6: Real consciousness is an awareness of a self, i.e. something that you identify with, that is beyond the realm of sensible experience, an awareness consistent with a simultaneous awareness of your self as the subject and object of spatiotemporal experiences.

§3. Feeling-Towards as a Model for Explaining Real Consciousness

We now have in hand a set of commitments that must be accounted for in explaining real consciousness. Two are about its content. **C1:** The content of the state—its intentional object—is the real self. **C2:** The content of the state is genuine freedom. Two have to do with its role in Kant's theorizing and account of the mind. **C3:** Real consciousness can be explained as a mental state with an intentional object that the haver of the mental state does not have a direct representation of. **C4:** Real consciousness is a third, conceptually distinct kind of consciousness. And two are about the context and significance of the awareness. **C5:** The explanation of real consciousness is neutral, i.e., it can be properly integrated into both its theoretical and practical contexts without

²⁴ When I say 'distinct' here, I mean that the self that is the referent of real consciousness is a distinct referent from the empirical self. I am not committing to saying that real consciousness gives us awareness of a distinct individual. Purely from considerations of the nature of consciousness itself, I contend that we do not get individuality. Each of our real selves, for all we are aware of, could be (e.g.) one part of a Spinozistic whole. However, Kant's moral commitments clearly do rely on noumenal individuality. The full account of the Kantian self, then, can wheel in these commitments in order to defend the idea that the self that we are aware of in real consciousness is an individual.

presupposing or exclude the other. **C6:** Real consciousness is an awareness of a self, i.e. something that you identify with, that is beyond the realm of sensible experience and consistent with a simultaneous awareness of your self as the subject and object of spatiotemporal experiences.

The obstacles to giving an account of real self-consciousness track the ones that stand in the way of integrating the real self into Kant's account of personal identity. These commitments are construed as individually falling afoul of Kant's commitments concerning the epistemic limitations and our knowledge of metaphysical truths discussed in §1. Therefore, they are not taken seriously as part of Kant's account of consciousness or self-consciousness.²⁵ If these cannot all be integrated into one account, we are left to conclude that this 'consciousness' is a bit of dogmatism or an assertion of mystery: unargued for, undefined, and undeserving of serious comparison and identification with the cognitive relation picked out by consciousness proper.

§3.1. Feeling the way to Real Consciousness: Identifying an Analogous Case

I advance a theory of real consciousness that satisfies all these commitments. The first step is to understand real consciousness as analogous to the feeling towards, like that in the case of feeling fear. Based on this, I argue that we can explain and characterize real consciousness at face value.

We can do justice to the identified commitments for explaining this state and to the work this

²⁵ Often, the claims Kant makes about consciousness of the real self are unremarked upon or read away, especially from a perspective that holds that Kant's main or proper theory of consciousness and self-consciousness takes shape exclusively or primarily in the *Critique of Pure Reason*. When they are noted, they are typically dismissed as inconsistent or inchoate on the grounds that at least one of the commitments I've identified cannot be accounted for. Kraus holds that to take seriously C1 is to, borrowing a phrase from Kant, "hypostasize the idea": make a thought into a real thing (2020, 182, 188-9). Longuenesse takes it that C5 is inconsistent with Kant's main theory of consciousness, and in entertaining such a state Kant is "in the grip of his own paralogism" (2017, 152). Kitcher (1984, 123) and Peter Strawson (1966, 249) both claim that a commitment to C6 is "beyond the bounds of intelligibility". Even when real consciousness is taken seriously, it is presented as inexplicable: a brute, ad hoc state. For instance, Wuerth (2014, 183-5) recognizes that Kant is committed to the existence of a real self and asserts that this means we have an "awareness" of the thing in itself: he does not provide an explanation for *how* we can come to have this awareness in the face of Kant's other epistemic commitments.

consciousness does in Kant's system without sacrificing his claims about our epistemic access to things in themselves.

As a reminder, here is the analogy to feeling fear that informs my main argument.

You are walking alone in a forest and feel a sense of fear, but you do not locate the cause of your fear. You feel bodily sensations that you associate with fear: a quickened pulse, goosebumps, and rising hair on the back of your neck. A plausible way of construing this case is by saying that the referent of *your feeling of fear* is the external unperceived fearsome object. Your physical reaction is symptomatic of this feeling towards, but is not the object of fear itself. In this case, there are two distinct intentional feelings: a bodily feeling and a feelings-towards.²⁶

I take this case to show that we can individuate two distinct but intimately connected mental states: the bodily feeling is direct result of the feeling-towards, and indirectly a result of the object that prompted the feeling of fear. The experience of a bodily feeling indicates both that there *exists* a thing that is eliciting your fear and that you are *aware* of the existence of this apparently fearsome thing. That is, it entails both a metaphysical fact and a cognitive fact: a fact about the kinds of mental states an agent has and the existence of a thing to which she stands in a cognitive relation. For the bodily feeling to occur at all, the feeling-towards—the feeling of fear—must have an intentional object that is not directly represented by the agent.

²⁶ Goldie (2002) introduces this terminology as a way to distinguish what he takes to be two components of an emotion, and to argue that emotions are not reducible to bodily sensations. What I need for my case is the idea that in feeling fear, you take there to be describable referent of the fear that is above and beyond your physical reaction.

Note that you cannot be construed as *representing* the object of your fear even if you vividly imagine all the possible things that could be dangerous out in the woods. Your fear is directed towards that fearsome thing though you do not perceive it.²⁷ For even if you are coincidentally imagining an instance of the very thing that is causing your fear, you do not thereby have a representation of the actual thing that is making your hair stand up. Nevertheless, your feeling of fear is *about* an object. Absent your ability to represent it, this object is describable only as that it is fearsome. In experiencing your bodily sensations, you recognize that you have a feeling that is directed towards something that can be ascribed only the most minimal properties: *that* it exists, and *that* is causing you fear.

The main upshot is that this is a case where we attribute a mental state with an intentional object to an agent, even though that intentional object is not one that the agent is directly representing. I show that understanding real consciousness as analogous to this ‘feeling-towards’ demystifies the state while preserving its metaphysical consequences.

§3.2. Feeling the way to Real Consciousness: Kant’s Account

I argue that thinking about the nature of and occasion for ‘feeling-towards’ gives us a model for real consciousness. A good reason to take this model seriously is that we can find Kant grasping for a similar explanation.

²⁷ One way to make plausible that this is a state of affairs we could find ourselves in is by appeal to mechanisms like blindsight or infrasound. These are functions of our cognitive and sensory apparatus where our body or brain picks up on stimuli even though we are not consciously or knowingly aware of this, like when a tiger growls at a high enough frequency that you don’t recognize that you are perceiving it. It is a matter of current controversy in the philosophy of mind whether e.g. blindsight is actually unconscious or conscious perception. See Block and Phillips, 2016. All that matters for my analogy is that the agent is not consciously *representing* the object of blindsight, etc. The tiger’s growl causes me to feel fear (and have the corresponding bodily reaction to fear) despite my lack of representation of the tiger. I would conjecture, too, that this (a feeling of fear absent a representation of the cause) is not a foreign experience to most people: it seems to be what informs the folk psychological concept of a sense or an intuition of danger.

In a note in the *Prolegomena*, Kant claims that the *I* is not a concept because it is “nothing more than a feeling of an existence” (Prol. 4:334n).²⁸ The *Prolegomena* was written to clarify the first edition of the *Critique of Pure Reason*, wherein Kant refers to the ‘I’ as the semantic expression and representation of self-consciousness, indicating that he means for this “feeling of an existence” to clarify a kind of self-consciousness.²⁹ And that he was looking to the language of feeling to clarify that this consciousness is intentional but non-representational is tracked in Kant’s general account of feeling and sensation. Alix Cohen has recently offered a reading of emotions for Kant which relies on Kant offering an account of feelings as mental states with “borrowed intentionality”: that is, being about the subject and the subject’s activity, but then being directed by the subject to be about the world, parallel to the fear case.³⁰ Henry Allison and Hannah Ginsborg, relying on passages such as CJ 20:223, have defended views on which feelings are both intentional and non-cognitive.³¹

A further reason to take the analogy between consciousness and the fear-feeling case seriously is that it accords with how Kant thinks mental states are individuated. In his most rigorous

²⁸ The full note is as follows: “If the representation of apperception, the *I*, were a concept through which anything might be thought, it could then be used as a predicate for other things, or contain such predicates in itself. But it is nothing more than a feeling of an existence without the least concept, and is only a representation of that to which all thinking stands in relation (*relatione accidentis*).” One might worry that Kant here talks about representations, and so the object of this feeling (in this case, existence) is represented to us. Such a worry can be quickly appeased by the recognition that Kant must be eliding some steps here: the ‘I’ itself cannot be a feeling, and so we should take Kant to say that the ‘I’ and use of ‘I’ expresses a feeling. The ‘I’ is a semantic representation that expresses this feeling.

²⁹ Some readers (e.g. Proops 2021, 84-85) have taken that in describing what I call real consciousness as a kind of ‘feeling’, Kant is attempting to minimize the role of consciousness, which is otherwise (says Proops) “an epistemic state that has no official place in his epistemology”. Contrastively, my reading shows that describing self-consciousness as a kind of feeling is *illustrative* of just what kind of state this consciousness is, and *how* it fits into Kant’s epistemology.

³⁰ Cohen 2020, esp. 442. Her account also appeals to Goldie’s distinction as a contemporary parallel.

³¹ Allison 2001, esp. 51, 54; Ginsborg 2003. This view is in contrast to Guyer 1997, e.g. 103-5, who construes feelings as non-cognitive because they are brute sensations. If that is your preferred reading of feeling, then my reading (I take it) of consciousness as a feeling-towards is still plausible and comes with a great deal of explanatory benefits, it just loses some of the support that fitting it into Kant’s broader account offers.

taxonomy of mental states, the *Stufenleiter* (step-ladder) passage (A320/B376-7), Kant enumerates a number of mental states—sensations, perceptions, cognitions, intuitions, concepts, and ideas—and explains what distinguishes them. This explanation makes clear that a particular mental state is at least partly characterized by the features of the content to which it refers. This is an explanatory dependence: explaining the nature of the state relies on the explanation of its referent.³² Representations are distinguishable based on their content, and as the object of representation changes, so too does the representation. This is not just a particular change (as in, I have countably different representations, representation A, representation B, etc.), but a change in the kind of mental state that representation is. Content dependent, I have an idea-representation, or a perception-representation. The explanation of the representation depends on the object to which it refers. In the context of our argument, the upshot of this is that Kant's diagnosis of the fear case I've described should yield the same conclusion: that there are two distinct mental states present despite both contributing to one experience of fear. By Kant's own lights, if we can identify a distinct referent, we have a different mental state: since the referent of one feeling is a physical reaction and the referent of another is a fearsome object, the feelings are themselves distinct.

³² For example, a sensation is a representation which refers to a change in the subject of the representation. That this particular kind of representation is a sensation, and not some other kind of representation, is due to what that representation is *of*. This dependence is distinguished from metaphysical dependence, wherein the existence of the representation depends on the existence of its referent. This further kind of dependence may be a feature of representations, but answering the question of whether Kant thinks representations are really dependent on their referents and explaining other elements of the relationship between representations and their objects gets into questions about just what kind of an idealist Kant is, and the nature of transcendental idealism takes us too far afield here. See Allais 2005 (Part One) for a discussion of the various ways this idealism can be interpreted. In sum, we cannot infer solely from the relation of consciousness to a representation, or from the nature of a representation, that the object of that representation exists. See, e.g., A197/B242 for this point in Kant.

Finally, understanding real consciousness as analogous to a kind of feeling accords with an often-overlooked thread in Kant, namely the significance of *intellectual* feelings.³³ While Kant's considered view is that the faculty of feelings has as its domain sensibility and empirical facts, and that a *feeling* is a state that is caused by sensible representations (see CJ 20:206), he also talks about the moral feeling of *respect* as a feeling "produced solely by reason" (KpV 5:76) and compares this kind of recognition to our awareness of necessity (KpV 5:30). Making sense of real consciousness as a type of feeling becomes part of a difficult and interesting philosophical project of specifying and classifying different feelings.³⁴

Feeling talk is apt for securing a characterization of this real consciousness. This characterization takes seriously Kant's own analogy, it accords with his claims about feelings generally and shines a spotlight on Kant's project of classifying both sensible and intellectual feelings, and the particular analogy I've proposed is in keeping with the commitments of his theory of mental states. We have good reason to understand real consciousness as an analog to feeling-towards.

§4. How to be Conscious of Your Invisible Self

I now apply the fear-feeling case to real consciousness to offer an account of the state that satisfies the core commitments outlined.

³³ The majority of interpreters take any reference to 'feeling' to be a reference to a sensible experience. See, for example, Grenberg's (2013) account of how we access the moral law through feeling, where she takes that to be restricted to sensible feeling, and discussion of her argument in Ware (2015).

³⁴ This is a project that is picked up and developed at length in post-Kantian thinkers, especially Fichte and Schopenhauer, who notes—in a similar vein to my argument below—that awareness of particular acts of the will manifest as a feeling (see *The World as Will and Representation* Book 1 §21). In taking Kant to have a central role for a mental state characterized in terms of intellectual (non-sensible) feeling, I also bring to light this strong connection between his thinking about the self and accounts developed by later, Kantian-influenced thinkers.

The primary issue was that Kant's commitments entail that this consciousness is a mental state with an intentional "object", but one that we do not directly represent. Kant's claim in the *Prolegomena*, as I've spelled it out above, suggests that we gloss this kind of self-consciousness as a kind of feeling an agent has of their own existence. We can situate this feeling in our analogous case. In the fear-feeling case, the agent's feeling of fear has an intentional object, but they do not directly represent that referent of the feeling. The agent feels fear—is conscious of a fearsome "object"—without a direct representation of that object. This is because they recognize they feel fear based on the experience of certain bodily feelings. In experiencing the bodily feelings, they are aware that there is an unrepresented object that their fear-feeling refers to.

If we can map this analogy onto Kant's account of real consciousness, we can get a good grip on how Kant takes himself to be justified in appealing to a consciousness that directly refers to the real self, even though it cannot be accompanied by a direct representation of that self. In other words, reading real consciousness as analogous to a feeling-towards an unrepresented object will satisfy C3, offering a path for awareness of the reality of things as they are that does not violate Kant's epistemic strictures. §4.1 supplies the remaining pieces needed for a complete and successful application of the analogy. §4.2 discusses the explanatory power of this understanding of real consciousness for an account of the Kantian self.

§4.1. A Complete Model for Real Consciousness

In order for the application of this analogy to be successful, we require another step: we need an analog to bodily feelings. In the fear case, the bodily feelings are what make us aware that we are experiencing this further conscious awareness. If real consciousness is to be understood as a feeling-towards our own self, we must recognize it via a similar mechanism.

My main claim in this section is that the best candidate for the analog to bodily feelings is an agent's consciousness of performing certain free acts. This parallel already suggests itself from the passages we've canvassed, where Kant goes back and forth between *equating* consciousness of freedom and consciousness of the real self and claiming that consciousness of the self *follows* from consciousness of freedom. If we take awareness of our free activity to stand-in for bodily feelings in the analogous case, then the slide between equivalence and succession is exactly as you would expect. When we feel fear of an unknown object, the bodily sensations associated with fear and the feeling-towards the cause of our fear are two distinct episodes of awareness that constitute *one* experience. Likewise, the awareness of a free action and the awareness of your real self come together to form the experience of recognizing (in the sense of feeling described above) a self that is capable of performing free actions. They stand in a reciprocal relation, wherein awareness of free acts would not be possible without awareness of a self, and likewise awareness of these free acts is what indicates that we are conscious of this self. This close relation explains why Kant allows himself to equate them on occasion, even though properly speaking (and indeed in the majority of cases Kant describes) they are to be pulled apart.

This initial match is confirmed by two striking passages, one from the *Groundwork of the Metaphysics of Morals* and the other from *The Critique of Pure Reason* where Kant says that we are aware of actions that cannot have their origins in sensible experience, and this gives us awareness of ourselves as real.³⁵ The passages are good evidence because what we require for the successful

³⁵ Choi (2019) also points out and makes use of the similarity between the argument structure of the third section of the *Groundwork* and the B-edition of the *Transcendental Deduction*. She takes these arguments to be in service of a somewhat different conclusion than my own, viz. that there is a difference between the self in the "world of understanding" and

application of the analogy is evidence that Kant thinks we have *some* awareness of performing actions that cannot have their cause in sensibility; not direct *knowledge* or *cognition* of the reality of freedom itself or, therefore, a complete account of how Kant thinks of our epistemic access to our own freedom works. Both of the following passages assert that the human being is aware of some “pure activity” or “spontaneity” within themselves. The in-text numbering is my addition:

For, since [the human being] does not as it were create himself and does not get his concept a priori but empirically, it is natural that he can obtain information even about himself only through inner sense and so only through the appearance of his nature and the way in which his consciousness is affected – [1] although beyond this constitution of his own subject, made up of nothing but appearances, he must necessarily assume something else lying at their basis, namely his ego as it may be constituted in itself; and thus as regards mere perception and receptivity to sensations he must count himself as belonging to the *world of sense*, but with regard to what there may be of pure activity in him [2] (what reaches consciousness immediately and not through affection of the senses) he must count himself as belonging to the *intellectual world*, of which he has no further cognizance [3]. (GMM 4:451)

The I think expresses the act of determining my existence. The existence is thereby already given, but the way in which I am to determine it, i.e., the manifold that I am to posit in myself as belonging to it, is not yet thereby given. For that self-intuition is required, which is grounded in an *a priori* given form, i.e., time, which is sensible and belongs to the receptivity of the determinable [i]. Now I do not have yet another self-intuition, which would give the determining in me, of the spontaneity of which I am only conscious [ii], even before the act of determination, in the same way as time gives that which is to be determined, rather I merely represent the spontaneity of my thought, i.e., of the determining, and my existence always remains only sensibly determinable, i.e., determinable as the existence of an appearance. Yet this spontaneity is the reason I call myself an intelligence [iii]. (B157-8n)

Kant makes the same set of claims in both passages: a human being is in time, and so the only way that they can intuit themselves is as a sensible entity [1], [i]. However, there are certain actions that must originate outside of sensibility, as they are determining *of* sensibility, rather than

in the “intellectual world”, but we are in agreement that the *Groundwork* and the B Deduction are part of a continuous argument working out the nature of self-consciousness.

determinable (determined *by* sensibility) [2], [ii]. We are aware of these actions and insofar as we are aware of these actions, we are aware that there is a part of ourselves that is not in time [3], [iii]. Kant offers examples of these kinds of self-determined mental activities immediately before the two passages. In the *Groundwork*, Kant claims that “reason must regard itself as the author its principles independently of alien influences” (GMM 4:458). In the first *Critique*, he describes the understanding’s synthesizing activity as spontaneity because it must *precede* sensible experience, and so cannot originate in or from sensible experience.³⁶

In both instances, then, Kant clearly commits himself to the idea that we have some awareness of performing actions that do not have their cause in sensibility. There are certain actions that we perform that cannot have their origin in sensibility, either because their content extends beyond the possible content of sensibility (like the ideas of reason) or because their performance is a condition on sensible experience (like the synthesis of the understanding). They must be things that we really *do*: free actions brought forth from ourselves. Moreover, we are *aware*, albeit in a minimal sense, that these actions do not have their origin in sensibility. Taken together, we get the claim that we represent the effects of these activities in a way that indicates both that there is

³⁶ See B2, A77/B103, A114, B135, B152-3, and B162n. I take it that a full defense of this claim (that there are actions we perform that do not and cannot have their origin in sensible experience) comes from understanding the structure of mental faculties in Kant. In brief, mental faculties—among them the will, the understanding, and reason—are at least partially responsible for the structure of sensible experience: in particular, for the fact that nature has a law-like structure. Given this, these faculties stand in a metaphysical and explanatory priority relation over the relational order of nature: they cannot, at least in certain original activities, be subject to this order (they could subject themselves to this order, but that would make that act of self-subjugation a self-determined action, and so again make it one that cannot and does not have its origin in sensibility). As all that my argument requires here is minimal evidence that we are aware of free actions, the full details of this argument are beyond the scope of this chapter. I bring this up here only to illustrate the fact that there is a preponderance of arguments and evidence that we can muster to secure the claim that there are certain acts that we perform—acts of the will and acts of thinking—that we recognize as originating outside of sensibility and so as free, even if we cannot thereby have knowledge of the reality of freedom itself. For a different but related argument for why faculties must be transcendently real, see Jauernig (2021), p.279-80.

a non-natural cause or ground of these effects, and that we are conscious of this ground even though we do not represent it. This is exactly what we need for the successful application of the analogy: the analog for bodily feelings in the fear-feeling case is an agent's consciousness of performing certain free acts.

This parallel also helps secure that there must be a referent of your real consciousness: analogous to the fearsome object (i.e., an object that causes bodily feelings of fear) there is a *self* that causes consciousness of performing certain free acts. What we require in addition to the parallel is an explanation of what licenses my identification with the ground of my free actions: why, in other words, is it consciousness of a *self* and not of some other metaphysical ground. Kant's answer to this question, suggested by the passages above, exploits both the nature of free activity and the first-personal access that we have to such activities. The idea is that if I were not performing these activities, then I could not be *aware* of these activities in the way that I am which crucially includes an awareness of them as free. For to be aware of a free activity in the relevant way is to be aware of it as self-determined and therefore as *my* activity: it is, in Kant's words, because of someone's awareness of what "pure activity *in him* (emphasis added)" that "he must count himself as belonging to the *intellectual world*" and similarly because of my awareness of my spontaneity that "I call *myself* an intelligence (emphasis added)". That is to say, to be aware of pure, i.e., free activity in the relevant way is to be aware of that activity as *mine*. Importantly, this consciousness

alone does not entail *individuality*, viz. that the content of your awareness is one singular self, and not, e.g., a self that is one part of a collective whole.³⁷

Taken together with the parallel between bodily feelings and an agent's consciousness of performing certain free acts this means the following: just like you are aware that there is a fearsome object, i.e., that there is something causing your feelings of fear, when you feel bodily feelings of fear (even though you do not know anything further about this unperceived object, viz. whether it is a tiger, an opossum, or just rustling leaves), so too your capacity to produce actions that do not have their origin in sensibility guarantees that there is a *self* that grounds these actions (even if you do not know any further properties of this self). This model, then, explains why Kant is confident that there is no reference failure in the case of real consciousness. The corresponding object of the state of consciousness must be such that it is causing *my* thoughts and *my* actions to occur: it must be, Kant argues, *me*.³⁸

Let us take stock. I have offered a model for understanding real consciousness. Real consciousness should be understood as a specific kind of feeling, namely a feeling-towards some thing that is eliciting an action, but that you do not perceive or directly represent. In virtue of performing free

³⁷ Kant clearly holds that we *are* individuals at the level of metaphysical reality, but the argument for this individuality emerges from his account of the nature of and requirements on moral constraint (like, for instance, KpV 5:105). This individuality, then, follows from an argument about the demands of morality, not from consciousness alone. See n.24.

³⁸ Schafer (2023) offers a stronger reading of this claim: that our consciousness that we are free is part of what *makes* us free, and so in being aware of our freedom we are actively constituting our self as a free entity. This amounts to, argues Schafer, a kind of *maker's knowledge* of our selves in a similar way to the knowledge the intuitive intellect has of real things (p. 88, 98). My argument here is amenable to being strengthened in the way suggested by Schafer. This stronger version of the argument offers an even clearer safeguard against a possible misidentification objection.

acts of thinking, an agent is made aware that she is conscious of her self as the real ground of these actions, even though that self is not represented as such.³⁹

Mapping the experience of the bodily feeling of fear and the feeling-towards a fearsome object to the consciousness of freedom and the consciousness of a real self accommodates what I have called commitments **C1** and **C2**: that the content of the state—its intentional object—is the real self and that content of the state has something to do with freedom. Real consciousness has as its intentional object the real self, but the experience that includes that awareness involves a consciousness of free activities. Crucially, this way of thinking about real self-consciousness makes sense of *real consciousness* as an epistemic state distinct from sensible or intellectual representation, and so allows for appeals to real consciousness to be consistent with Kant's overarching claims about our epistemic and cognitive limitations. My account of real consciousness as a feeling-towards accordingly satisfies **C3**: real consciousness can be explained as a mental state with an intentional object that the haver of the mental state does not have a direct representation of. In explaining how we are and can be aware of performing free actions, we have also secured the neutrality of an account of real consciousness and so satisfied **C5**. We are conscious of our real self in both theoretical and moral contexts, although the introduction of a moral context does introduce an important disanalogy insofar as recognition of the moral law affords us an argument for individuality at the level of things in themselves.

³⁹ It follows from my reading that consciousness is an epistemically distinct and significant state. It is more robust than *thinkability*, as I can “think whatever I like with no assurance that there is a corresponding object” (Bxxvi). Neither, though, is it equivalent to either knowledge or cognition: states of significant comprehensive or understanding that require, at minimum, some determinate content. For more on Kant's technical notion of cognition and a debate about the determinate content requirement, see Watkins and Willaschek (2017) and Schafer (2022).

§4.2. *The Explanatory Power of the Feeling Model of Real Consciousness*

I've argued that real consciousness is a mental state that refers to an object that the agent does not represent as such, despite being aware that she is conscious of an object that she is not (and, in the case of the real self, could not) represent as such. I've argued, too, that the recognition of such a state is only possible because the performance of certain activities necessitate that we identify with the entity that could perform these actions, and so conceive of that entity as my *self*.

This way of understanding real consciousness offers an interesting Kantian answer to the problem of personal identity, an answer that I hold has some strong intuitive pull.⁴⁰ In describing this answer, I explain how Kant's account of real consciousness satisfies C6, and explain how we can have an awareness of a non-spatio or temporal self that is consistent with a simultaneous awareness of your self as the subject and object of spatiotemporal experiences.

If someone were to ask you *who you are*, you might first appeal first to relational facts: that you are a sibling, or a citizen of a particular country, or work a particular job. If pressed, you would not want to say that these facts constitute who you are. Insofar as they are relational, they are contingent, and so they cannot exhaust your description: all of these relations could have been other than they are. If *you* think *you* could work a different job, have been born in a different country, etc., then you must be tracking that there is some feature of *you* that is not exhausted by these relational facts. To fill this gap, one might be tempted to appeal to character traits: I am funny, or I am kind. But these character traits won't resist this pressure either: someone being funny is an entirely generic description absent particular experiences in which this trait manifests.

⁴⁰ The account I offer here is similar to one proposed by Wood (1984) and Watkins (2005, see especially p. 324) in the service of defending the metaphysics of Kant's account of freedom and causality.

Again, we either revert back to relational features and facts and run back into the same problem of characterizing ‘you’, or we say that the description of *me* is generic. That I am funny, kind, etc. could pick out *some* person, but they don’t specifically pick out *me*.

To be clear, I am not arguing that Kant appeals to a real self because it is not possible to conceive of counterfactual claims about relational identity facts. What I do take to be motivating Kant is the search for an *intrinsic* feature: something absolutely intrinsic that cannot be reduced to relational properties, and so in virtue of which our relational properties are explained.⁴¹

In isolating such an intrinsic feature, what we are left with is an awareness of a self that you can ascribe no properties to. It seems, in other words, that you are left with precisely the kind of state that Kant is cashing out in his description of real consciousness. On the account I’ve provided, I take it that his answer is one with intuitive force: that the self that we are trying to characterize can never be represented as such even though we are conscious that it exists. Our real self, whatever that may be, is like that thing in the bushes that causes your hairs to stand up. It is included in the referent of ‘my self’, or ‘I’: when you refer to yourself you are referring to the real self alongside your experiences, just like when you say ‘I feel fear’ you can plausibly be taken to refer to the object of your fear alongside your bodily experiences.

The real self, then, is something which we are aware of but that we cannot directly represent or ascribe concrete properties or descriptors to. Any attempts to do so will have to rely on experienced (spatiotemporal, relational) properties, and so fail to capture the exact object that we

⁴¹ One way to read Kant’s account of the relationship between the intelligible and the empirical characters in the Resolution to the Third Antinomy is as of such an intrinsic and relational relation. For further discussions of the relationship between the intrinsic and the relational in Kant’s idealism, see Allais (2005, chapter 9), and Hogan (2021).

are conscious of. This is, I take it, precisely the point that Kant is trying to articulate in the continuation of the passage from the *Groundwork* we saw above. He notes that we are conscious of our real self, and then asks what it would be to attribute properties to this self.

[...] but with regard to what there may be of pure activity in him (what reaches consciousness immediately and not through affection of the senses) he must count himself as belonging to the *intellectual world*, of which he has no further cognizance. A reflective human being must come to a conclusion of this kind about all the things that present themselves to him; presumably it is also to be found even in the most common understanding, which, as is well known, is very much inclined to expect behind the objects of the senses something else invisible and active of itself — but it spoils this again by quickly making this invisible something sensible in turn, that is, wanting to make it an object of intuition, so that it does not thereby become any the wiser. (GMM 4:451-2)⁴²

I take Kant to be arguing that, although there *is* a real existence that I *am* conscious of, projects that seek to describe this existence in any greater detail must by necessity revert to the properties, predicates, or concepts that come out of the sensible world. By attempting such a project, one thereby inadvertently end up attempting to describe this self as object in the world.⁴³ In so doing, one (often unwittingly) slips into describing a different object, and so fails to capture the thing they are trying to describe.

Kant is taking seriously that we have an intrinsic sense of self, to the point of substantiating it with an underlying metaphysics. It is a commitment to this metaphysical picture that allows him to realize what I take to be a fruitful account of the cognitive relation in which you stand to that self. On this understanding, we can preserve our everyday recognition of how the world around me and my relation to it features in my representation of my identity while still maintaining,

⁴² See also B431-2 for an almost identical set of claims, as well as B156 and Kant's argument there that you cannot help but represent your self as an object analogous to the way that one cannot help but represent time as linear.

⁴³ This, I take it, is the problem at the heart of the Paralogisms: the rational psychologist thinks that they can have an *intellectual, determinate representation* of their self, but they are really confusing the application of concepts that can only properly be applied to sensible experience.

following Kant, that a conception of my identity is incomplete without reference to this sense of a real self despite the minimal information it yields.

I have now explained what it is to be conscious of your real self, for Kant. On my view, real consciousness is analogous to a ‘feeling-towards’ where you do not represent the referent of that feeling, only the states that it elicits for you. In the case of real consciousness, the elicited ‘states’ are free acts of the mind, activities which are metaphysically supported (caused or grounded) by the existence of a real self. I’ve argued, too, that such a consciousness of a *self* (not merely some entity) unindexed to spatiotemporal experiences tracks an intuitive feature of the phenomenology of self-awareness and identity: that you are aware of something intrinsic about your self, but you cannot give a further characterization of this intrinsic thing because it, by virtue of what it is, cannot be described with relational properties. While many philosophers—perhaps most notably, Kant’s predecessor, Hume—maintain that we should be skeptical about such an intrinsic feature or give an error theory for it⁴⁴, it is fruitful for our interpretation of Kant to recognize that he takes his metaphysics of the self and morality as part of a project to substantiate this feature.⁴⁵

§5. Conclusion

I began with Kant’s surprising claim that we are conscious of our invisible self. I have argued that this claim should be understood as part of an account of a specific kind of self-consciousness: real consciousness, the consciousness that we have of ourselves as metaphysically real entities or

⁴⁴ See Hume’s famous claim in the *Treatise of Human Nature* that there is no abiding self because there is one impression of the self (*Treatise*, 1.4.6). For a contemporary defense of the self as a set of relational experiences, see Ismael (2006).

⁴⁵ I take it that this is fruitful both for how we interpret Kant’s theory, *and* fruitful for the philosophical consequences and applications that we can take the theory to have. I discuss one such fruitful application in the dissertation’s conclusion, drawing on both the account of a feeling-towards your self that I develop here, and the way I spell out the metaphysics of this theory (discussed in detail in chap. 3, esp. §4).

things-in-themselves. I explained this self-consciousness as analogous to a feeling that we have towards our real existence, a feeling that we recognize in virtue of activities that do not have their origin in sensibility. We really are conscious of the invisible self: we feel it, even if we do not represent it.

In so explaining, I have opened the space to offer a consistent and comprehensive account of Kant's theory of personal identity. He takes his own answer to the question 'what makes the person that I am?' to rely on an appeal to the self as it is in itself, but in order for such an answer to be possible there needs to be an explicable cognitive relation that an agent can stand in to their real self. I have shown that we do find such a relation in Kant's theory. While providing an account of this relation—real consciousness—is not equivalent to offering a full account of personal identity in Kant, an ability to appeal to such a relation removes the main barrier to integrating all of Kant's claims about the real self and the person into one systematic theory.

Removing this barrier was my main goal here. This work is not yet entirely done. I've argued that understanding real consciousness as a feeling-towards allows us make sense of it as a mental state consistent with Kant's idealist commitments. I hold, too, that understanding its properties as such reveals that Kant appeals to *three* distinct kinds of self-consciousness. While my positive account has done some work to motivate this claim about distinctness, I have not yet fully defended this claim. To invoke the commitments I introduced here, I have not yet mounted a full explanation and defense of **C4**: Real consciousness is a third, conceptually distinct kind of consciousness. I turn to offer this now, in chapter 2.

Chapter 2

Kant's Three Kinds of Consciousness

How many kinds of consciousness does Kant think we have? Scholarly consensus is that he describes and makes use of *two* distinct kinds of self-consciousness. This consensus is rarely explicitly stated as such, but it is driving assumption across a myriad of work on Kant's theory of consciousness, the mind, and the self. This assumption is mentioned almost in passing in several seminal writings on Kant's overarching theory (e.g. in Karl Ameriks (2005, 142), Paul Guyer (1987, 83-85), and Henry Allison (2004, 168)), it is taken as the fixed starting point for further investigation (by, e.g., Janum Sethi (2021) and Olga Lenczewska (2019)), and situated as the inspiration for contemporary bipartite views of consciousness (see, e.g., Tyler Burge (1997) and Matthew Boyle (2009)).⁴⁶ The rare view that challenges this consensus—like that of Béatrice Longuenesse (2017, 86) and Heinz Heimsoeth (1967)—does so indicating that is a deviation from the norm to suggest that Kant entertains any more than two kinds of self-consciousness.⁴⁷

⁴⁶ Allison revises this view somewhat in (2015) as discussed below in n.48, although he ultimately seems to conclude that Kant has two kinds of self-consciousness. Likewise, Sethi suggests at one point that there may be a principled distinction between pure and transcendental apperception, but does not pursue this further in the paper. The authors mentioned here far from exhaust the appeals to Kant's two kinds of self-consciousness: see also, e.g., McLear 2020, Proops 2021, Wuerth 2014, Merritt 2018, Henrich 1982, 1994, Indregard 2018, Longuenesse 2023. Further, Kant is often credited with being one of the first to distinguish between (what we now call) object-consciousness and subject-consciousness: that is, awareness of myself as an object (as in, 'my arm is broken') and awareness of myself as the subject of experiences (as in, 'I am in pain'). For more discussion of these kinds, see Wittgenstein 1958, Evans 1982 215-20, Longuenesse 2017 (chapter 1).

⁴⁷ Heimsoeth 1984 mounts the more explicit challenge of this consensus, though he does not spell out at any length how we should understand each of the three kinds. Longuenesse does say that Kant describes a *third kind* of consciousness of thinking (2017, 86), but it is unclear if she ultimately thinks this amounts to a third, distinct kind of self-consciousness. Ginsborg 2013 is another scholar who challenges this consensus, but she does so by denying that there is any significant difference in the content of empirical and transcendental consciousness (viz., on her view, there is really *one* kind of self-consciousness). Kraus (2020, 129) offers a taxonomy of self-consciousness that includes *five* different types, but—as I discuss at greater length in §2 and §3—it is ultimately unclear what the distinction between several of these types actually amounts to, and so I argue she can plausibly be read as committed to these two kinds of self-consciousness, roughly consciousness of the self as subject of thought, and consciousness of the self as object.

Excepting these rare instances, the question is taken to be settled (perhaps before it was even properly asked): Kant appeals to two kinds of consciousness. Call this the Two Kinds View.

Here is one way to generate worries, or at least suspicions, about the Two Kinds View. It is true that Kant will frequently refer to pair-concepts of consciousness: that is, he will contrast something as ‘consciousness type 1’ to ‘consciousness type 2’. But what ‘type 1’ and ‘type 2’ are specified as varies widely across passages and across texts. Frequently, Kant distinguishes between “pure” consciousness and “empirical” consciousness, as he does at A116, B132, and Anth. 7:135n. Sometimes, the distinction is drawn as between *inner sense* and *transcendental* consciousness or apperception. Occasionally—but not always—inner sense and empirical consciousness are apparently equated, and occasionally—but not always—transcendental and pure apperception are apparently equated.⁴⁸ In other instances (G 4:457, KpV 5:162), Kant contrasts “consciousness of myself as an object affected through the senses” with “consciousness of [myself] as an intelligence”. Rare moments (see A546/B574) see these various concepts and references associated, but how exactly we should understand the relations and equivalence or lack thereof between these various concepts is not evident. All of these different pairs are appealed to across different versions of a Two Kinds View, with little consensus about which, if any, are definitive.

Given this proliferation of concepts and concept-pairings and given the tension between the apparent consensus *that* Kant describes two kinds of consciousness with the lack of clarity on

⁴⁸ I go into the details of the different types of apperception Kant refers to in §1. The difficulty of straightforwardly equating ‘inner sense’ and ‘empirical consciousness’ is pointed out by Allison (2015, 357-9).

what these two kinds are or consist in, we should at least be suspicious that Kant is straightforwardly or clearly committed to there being two kinds of consciousness. Surveying these passages is not sufficient to mount an argument that Kant explicates more than two kinds of consciousness. They do, though, open the door to revisiting the question about kinds of consciousness in Kant's account. That revision is what I offer in this chapter.⁴⁹

Since the text is not clearly decisive on the question of kinds of consciousness, we should seek a criterion for identifying a kind of consciousness and employ this criterion to generate an account of the number of kinds. Here is a simple, but more rigorous, way to demarcate kinds of consciousness in Kant's theory: we can identify a distinct kind of consciousness if its primary properties are incompatible with the primary properties of a previously identified kind of consciousness.

This criterion, I argue, is what is in the background of the Two Kinds View. I spell out why that is the case in §1. But identifying and employing this criterion reveals, I argue, that Kant is committed to three conceptually distinct kinds of consciousness. This argument takes up the bulk of the paper. My argumentative strategy in §3 and §4 begins by identifying a kind of consciousness referred to in several key arguments at the heart of Kant's main discussions of self-consciousness and the self. These arguments are all notoriously difficult to understand, and I argue that this difficulty is the result of an attempt to integrate the properties and function of that kind of consciousness into either of two previously identified kinds of consciousness. The result

⁴⁹ In making my argument here, I go back and forth between referring to kinds of *consciousness* and kinds of *self-consciousness*. This is both for ease of reference and because I take it that each kind of consciousness in Kant has a corresponding self-consciousness and vice versa: each kind of self-consciousness belongs to a general kind.

of this integration is a state with incompatible or contradictory properties, thereby rendering Kant's arguments opaque if not nonsensical. To avoid this undesirable result, we should read that kind of consciousness as a third, distinct kind. Since the clarity and comprehensibility of several of Kant's arguments hang on there being three kinds of consciousness, I conclude that there are three, not two, kinds of consciousness in Kant: what I call the Three Kinds View.

This conclusion and my defense of the Three Kinds View picks up on and supplements my main argument in chapter 1. I there identified an interpretative commitment I labeled as *Claim 2*: Kant's account of self-consciousness is limited to two kinds: empirical self-consciousness, and formal self-consciousness. I argued that theories that are committed to only two kinds of consciousness are thereby committed to theories of personal identity that cannot accommodate all of Kant's claims about identity and the self. I claimed that we need to recognize a different kind of consciousness that can bridge the gap, and I argued for this consciousness and its place in Kant's theory. Insofar as I did this, I have argued indirectly for the Three Kinds View, where I take it that what I have called 'real consciousness'—the consciousness that we have of our metaphysically real self—is the distinct, third kind of consciousness in Kant's theory. But, as it is stands, I have not yet argued specifically that real consciousness is a *third kind*. I have not defended the commitment of real consciousness that I there labeled **C4**: Real consciousness is a third, conceptually distinct kind of consciousness.

Without a defense of this commitment, my argument in chapter 1 is open to an objection of double-counting: that I am attributing a distinct kind of consciousness to Kant when I should be integrating the references to consciousness of a real self into the account of either empirical or

formal consciousness. In arguing that Kant's arguments require an appeal to a third kind of consciousness, I will substantiate and secure my claim that real consciousness is this distinct kind.

§1. Two Kinds of Consciousness

I begin with a refresher on the two kinds of consciousness that I introduced in chapter 1 and that I there argued informed *Claim 2*. Here, I will go over the basics of these two kinds—empirical consciousness, and formal consciousness—in order to argue that (a) Kant is clearly committed to there being these two kinds, and (b) these two kinds are conceptually distinct. I will use this to identify our criterion for demarcating them as distinct kinds.

Up to this point, I have been speaking somewhat vaguely of the 'two kinds of consciousness' that underly the Two Kinds View. Some of this vagueness is for good reason. In part, it is motivated by the considerations I raised in introducing the question about the different kinds of consciousness that Kant makes use of, viz. that he uses a range of different terminology without clear indication when and where these terms are co-extensive. Another part has to do with the variety of ways one could spell out the details of the Two Kinds View: I turn to consider three different options in the next section. Nevertheless, with these considerations in back of mind we can here broadly settle the basics of two kinds of consciousness: what they are, and why they should be distinguished.

I take the two kinds of consciousness that underly the Two Kinds View generally construed are just the two kinds that I identified in the previous chapter: empirical consciousness, and formal consciousness. Empirical consciousness is the consciousness that attaches itself to particular, spatio-temporal representations. Formal consciousness is the awareness of one's self as the subject

of representation: that is, as the underlying or formal condition for the experience of all *your* particular spatio-temporal representations.

There is plenty of textual evidence that Kant is committed to these two kinds and committed to them *as* two distinct kinds. For instance, in the following passages he clearly describes one consciousness, viz., empirical consciousness, as an awareness of discrete and varied empirical objects and perceptions, and then points to *another*, separate consciousness that is a necessary condition for the possibility of these empirical perceptions, viz. what I have been calling formal consciousness and what is here referred to as transcendental apperception.

Now this original and transcendental condition is nothing other than the **transcendental apperception**. The consciousness of oneself in accordance with the fluctuations of our state in internal perception is merely empirical, forever variable; it can provide no standing or abiding self in this stream of inner appearances, and is customarily called **inner sense** or **empirical apperception**. That which should **necessarily** be represented as numerically identical cannot be thought of as such through empirical data. There must be a condition that precedes all experience and makes the latter itself possible, which should make such a transcendental presupposition valid. (A107)

One should attend carefully to this proposition, which is of great importance. All representations have a necessary relation to a possible empirical consciousness: for if they did not have this, and if it were entirely impossible to become conscious of them, that would be as much as to say that they did not exist at all. All empirical consciousness, however, has a necessary relation to a transcendental consciousness (preceding all particular experience), namely the consciousness of myself, as original apperception. (A117n)

Thus all manifold of intuition has a necessary relation to the I think in the same subject in which this manifold is to be encountered. But this representation is an act of spontaneity, i.e., it cannot be regarded as belonging to sensibility. I call it the pure apperception, in order to distinguish it from the empirical one, or also the original apperception, since it is that self-consciousness which, because it produces the representation I think, which must be able to accompany all others and which in all consciousness is one and the same, cannot be accompanied by any further representation. (B132)

Here is the reasoning behind my choice of terminology. While I follow Kant in my use of ‘empirical consciousness’, I adopt the expression ‘formal consciousness’ to both consolidate a

range of different expressions and terminology and to differentiate from a set of other terms. The more familiar (at least, to readers of Kant) choice is the expression employed by Kant in the above texts, namely, transcendental apperception. However, I contend that without a quite detailed clarification, this term is simultaneously unhelpfully loaded and unhelpfully underspecified.

Both <apperception> and <transcendental> come with baggage that stands in the way of our understanding. Kant is very consistent with how he deploys 'empirical' apperception. Beyond this category, though, he refers to 'transcendental' apperception, 'mere' apperception, 'pure' apperception, and 'original' apperception.⁵⁰ There is some reason to take some of these terms as equivalent, other reasons to think that they are distinguishable, and it is certainly not clear cut whether 'transcendental apperception' is the very same thing as 'original', or 'pure', or 'mere' apperception. I here begin the project of properly cataloguing these kinds and their relations here. But at the outset of such a project, an appeal to non-empirical apperception without specification confuses the investigation.

The same holds true for <transcendental>. While I do not deny that apperception is very importantly transcendental and we do not want to lose sight of this, 'transcendental' in Kant is often ambiguous between (a) a necessary condition for the possibility of experience, and (b) an object that is real and exists outside of possible experience. There is of course room to clarify this, but that clarification is itself lengthy, contentious, and relies on conclusions that have yet to be

⁵⁰ See A94/B127, A108, A111, A116, A123, B133, A343/B401, A400, B546/B574 for the range of different terms. And this is not yet to say anything about the varieties of *unity* of apperception: original unity, transcendental unity, synthetic unity, analytic unity, absolute unity. It is evident (see, e.g., B132) that the unity or unities of apperception are something above and beyond apperception itself, but how apperception composes unities and what relations it stands in to what in order to do so are questions that can only be answered upon a more thorough taxonomy of the kinds listed here.

established. For the purposes of *undertaking* an investigation into Kant's concepts, the presence of that ambiguity threatens to confuse the undertaking before it begins.

What I aim to do with the expression 'formal consciousness' is shift the debate away from terminological ambiguity and focus instead on the key features that Kant is tracking when he refers to the kind of apperception or consciousness that is a condition on the possibility of thought and further representations. Why 'formal' consciousness, then? Because this picks up on a consistent descriptor that Kant relies on to explicate the features of this kind of consciousness. Here is one example:

At the ground of this doctrine we can place nothing but the simple and in content for itself wholly empty representation I, of which one cannot even say that it is a concept, but a mere consciousness that accompanies every concept. Through this I, or He, or It (the thing), which thinks, nothing further is represented than a transcendental subject of thoughts = x , which is recognized only through the thoughts that are its predicates, and about which, in abstraction, we can never have even the least concept; because of which we therefore turn in a constant circle, since we must already avail ourselves of the representation of it at all times in order to judge anything about it; we cannot separate ourselves from this inconvenience, because the consciousness in itself is not even a representation distinguishing a particular object, but rather a form of representation in general, insofar as it is to be called a cognition; for of it alone can I say that through it I think anything. (A346/B404)⁵¹

This passage illustrates that this consciousness, whatever we name it, represents nothing further than the subject of thought, and to represent the subject of thought is just to be the formal condition on, or the form of, representation. Hence, formal consciousness.

With this in hand, let me return to the distinction between the two kinds. As I said, Kant clearly intends to identify them as distinct. And further, we can reify this distinction with principled

⁵¹ See too A398: the logical unity of thought is the "formal condition" on thinking; A321/B378: the "mere logical form of our cognition can contain the origin of pure concepts"; A354: the I think is the "form of apperception".

criteria. Part of the reason it is worth doing the ground-clearing work on how to refer to the state of formal consciousness is that it helps to home in on the state's primary or defining properties. The key features that Kant is tracking with this kind of apperception/consciousness is (a) its structuring role, and (b) the lack of inherent content to this structure (its emptiness). Formal consciousness is the awareness of the subject of representation and so it is formal: it does not have representational content. Empirical consciousness, on the other hand, is the awareness of particular representational content. The kinds are demarcated by having distinct and non-overlapping properties: one has particular, spatiotemporal content, the other has no content.⁵² They are further distinguished by virtue of standing in an asymmetric dependence relation: the possibility of empirical consciousness (awareness of particular complex representations) depends on formal consciousness as the unified structure that makes possible these complex representations (see, for instance, B133).

We can generalize from the specific distinction between empirical and formal consciousness to a criterion for identifying when a state of awareness counts as an instance of a distinct kind of consciousness. Here is that criterion: we can identify a distinct kind of consciousness if its primary properties are incompatible with the primary properties of a previously identified kind of consciousness.

⁵² Specifying this content/no-content distinction in greater detail results in further tensions or contradictions between the properties of the two states: the content of empirical consciousness is varied, without consistent identity, and contingent. The structure of formal consciousness is consistent, identical throughout, and necessary.

I will employ this criterion to argue that Kant is committed to a further, third distinct kind of consciousness in addition to empirical and to formal consciousness. I now turn to the kinds of views and commitments I am arguing against in making this claim.

§2. Three Different Paths for the Two Kinds View

I've argued that Kant's view of the structure and content of thought results in a commitment to two distinct kinds of consciousness. I used the way these kinds are distinguished to generate a criterion for identifying a distinct kind of consciousness in Kant. According to this criterion, we can identify a distinct kind of consciousness if its primary properties are incompatible with the primary properties of previously identified kind of consciousness.

My main argument in the paper is to employ this criterion to identify a third kind of consciousness in Kant. This section will motivate the need for such an argument. I offer a sketch of a kind of consciousness whose form and content seem *prima facie* at odds with the primary properties I've identified of empirical and formal consciousness. In the face of these references, though, the response in the literature has been to double down on the Two Kinds View by integrating these remarks into the concept of one of the two previously identified kinds of consciousness. I look at three different strategies for integration readings.

§2.1. Consciousness X: A Puzzle Case

Here is a refresher on my argumentative strategy. The argument in this chapter provides both further support and abductive evidence for my view about real consciousness as articulated in chapter 1. It does so by arguing that we can conclude to real consciousness via a different route, viz. by identifying a gap in Kant's taxonomy and concepts of consciousness. What I aim to show here, then, is that the content of the consciousness that I identify in this sub-section cannot be

explained as a part of the content or function of one of our two existing kinds, giving good reason to conclude that it is conceptually distinct. To successfully make that argument, I don't want to presuppose the conclusion. For that reason, while I will here refer to several of the same passages and considerations that I used to motivate my account of real consciousness, I ask the reader to bracket that account for the time being. I will show here that, even without that account in hand, we can arrive at the need for a distinct state of consciousness that has the features I've attributed to real consciousness.

The following passages all refer to some kind of consciousness [*Bewußtsein*]. The bold emphasis is my addition.

1. But the very same subject, being on the other side **conscious of himself as a thing in itself**, also views his existence insofar as it does not stand under conditions of time and himself as determinable only through laws that he gives himself by reason; and in this existence of his nothing is, for him, antecedent to the determination of his will, but every action – and in general every determination of his existence changing conformably with inner sense, even the whole sequence of his existence as a sensible being – is to be regarded **in the consciousness of his intelligible existence** as nothing but the consequence and never as the determining ground of his causality as a noumenon. (KpV 5:97-8)
2. For, that a thing in appearance (belonging to the world of sense) is subject to certain laws from which **as a thing or a being in itself it is independent** contains not the least contradiction; that he must represent and think of himself in this twofold way, however, rests as regards the first on consciousness of himself as an object affected through the senses and as regards the **second on consciousness of himself as an intelligence, that is, as independent of sensible impressions** in the use of reason (hence as belonging to the world of understanding). (GM 4:457)
3. But suppose there subsequently turned up—**not in experience but in certain (not merely logical rules but) laws holding firm a priori and concerning our existence**—the occasion for presupposing ourselves to be legislative fully a priori in regard to our own existence, and as self-determining in this existence, then this would disclose a spontaneity through which our actuality is determinable without the need of conditions of empirical intuition, and here we would become aware that in the **consciousness of our existence something is contained a priori that can serve to determine our existence**, which is thoroughly determinable only sensibly, in regard to a certain inner faculty in relation to an intelligible world (obviously one only thought of). (B431-2)

4. Yet the human being, who is otherwise acquainted with the whole of nature through sense, **knows himself also through mere apperception**, and indeed in actions and inner determinations which cannot be accounted at all among impressions of sense; he obviously is in 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. (A546/B574)⁵³

Very similar language is used in each case: intelligible' or 'intelligence'; self-determination; *a priori* determination. This consistency gives a good first-pass reason to hold that the same concept of consciousness is being invoked throughout these discussions. Call the state in question Consciousness X.

The question at issue here is whether Consciousness X is empirical consciousness, formal consciousness, or a third distinct kind. Prima facie, the properties and content attributed to Consciousness X are in tension with the properties of both formal and empirical consciousness. Consciousness X is described as having an *object*, or at least some minimal content, contrary to the awareness of a bare or empty subject attributed to formal consciousness. But, in contrast to the objects of empirical consciousness, that object or content is not particular, spatio-temporal content: it is of a thing in itself, something beyond the impressions and receptivity of the senses.

On the face of it, then, it seems that Consciousness X should be a distinct, third kind of consciousness. But there might be reason to question or reject this prima facie tension. As I noted at the outset, Kant frequently appeals to pair-concepts of consciousness (even if these pairs are not consistent). But these appeals are taken as good reason to want to maintain, if possible, a Two Kind View of consciousness on Kant's behalf. The sanctity of the pairs can be bolstered by an

⁵³ See also KpV 5:99, Anth 7:141n, GM 4:457, B157-8n. The appeals to this consciousness in the *Critique of Pure Reason* will be examined in detail in §3 and §4.

appeal to parsimony: if two kinds or concepts of consciousness can do all the work we need them to, then all things considered we should prefer that account over one that posits further kinds. Finally, Kant's epistemic humility thesis—that human minds cannot cognize or know things as they are in themselves—is traditionally interpreted to mean that we have no access *at all* to things in themselves. If this is the right way to read epistemic humility, then the passages I take to be in tension with formal and empirical consciousness must have a different interpretation, or else generate a contradiction with one of Kant's key doctrines.⁵⁴

Based on any of these considerations, one might strive to preserve the Two Kinds View. In light of these passages, the way to do so is to show that the properties of Consciousness X can be integrated into the concept of either formal or empirical consciousness. I take there to be three different interpretive options for such an integration, interpretations against which I develop my own argument. The options I have in mind are: (1) consciousness X is formal consciousness (deflated reading); (2) consciousness X is formal consciousness (inflated reading); and (3) consciousness x is empirical consciousness. In §2.2-2.4 I introduce each in turn by drawing on some views from recent literature.⁵⁵ Each option is given here and will be discussed and criticized in §3. In discussing views in the literature, I do not mean to imply that they fully instantiate the label I give them. Especially since this question is rarely considered, and when it is scholars have

⁵⁴ I addressed this worry in defending real consciousness in the previous chapter. As I conclude here that there is a third kind of consciousness and that that third kind has the properties of real consciousness, my arguments there responds to this objection. I raise it again here in order to bring to light why scholars have been reluctant to embrace a third kind of consciousness.

⁵⁵ In chapter 1, I focused on scholars who take the claim about two kinds of consciousness to rule out the inclusion of the real self in Kant's theory of the self and personal identity. Here, I go into greater length on the various interpretive paths that have been pursued in order to maintain a commitment to the Two Kinds view.

tended to embrace attributing (problematic) contradictions or tensions to Kant on this topic⁵⁶, few authors have straightforwardly subscribed to the readings I describe below. Nevertheless, I take it that each exhibit tendencies that fit into the options as described here.

§2.2. Option 1: Consciousness X is Deflated Formal Consciousness

The most popular option for a Two Kinds View is to read descriptions of Consciousness X as descriptions of formal consciousness (or transcendental apperception). This option, though, comes apart in two quite different interpretive strands which each deserve their own consideration. The first, and the one I consider here, I call ‘Deflated Formal Consciousness’: I’ll refer to its proponent as a deflater. The deflater holds that Consciousness X is formal consciousness. In order to make passages of the kind cited above fit to this thesis, they *deflate* the parts of the text where Kant appears to be speaking of consciousness of things in themselves. They reject the view that there is any real or substantive metaphysics in Kant’s account of the mind. In so doing, they take the description of Consciousness X to be compatible with the original definition and basic features of formal consciousness.

Versions of this view can be found in the majority of interpreters of Kantian consciousness, including in Béatrice Longuenesse⁵⁷, Peter Strawson, Andrew Brook, Ian Proops, and Tobias

⁵⁶ See, for example, Strawson: “there is no refuge but incoherence from the question how the connexion is to be made, in the way of identity, between the natural being, the man, with a mental history of thoughts, perceptions, and feelings and the supersensible being, with no history at all” (1966, 249); and Jacobi’s original criticism of the system of transcendental idealism that there is an inexplicable, “mystical” connection between the object and subject (1802, 1:269).

⁵⁷ Notably, I class Longuenesse under this option even though I have previously noted that she suggests that Kant has three kinds of consciousness. She identifies a *third kind* of consciousness of thinking in the part of the Deduction that is the focus of my §3, describing it as “the mere consciousness of the act of thinking” and “as close as Kant gets to admitting pure action awareness” (2017, 86). But though she identifies it as such she does not—as far as I can tell—take Kant’s account of the self to rely on any reference to this consciousness. This consciousness either does not reappear in her account, or it reappears under the guise of one of more familiar two kinds, implying that this kind of consciousness

Rosefeldt.⁵⁸ Katharina Kraus, whose interpretation is guided by consideration of the theoretical passages that refer to Consciousness X, characterizes the state that I refer to as formal consciousness as the specification of the *form of reflexivity*: it plays its structuring or formal role by virtue of relating a set of representations to a single source.⁵⁹ It is not an “assertion of existence” of any kind, but an “indeterminate expression” that “designates the subject of apperception, though without determining it as the content of a representation”.⁶⁰ Kraus holds that the references to Consciousness X are Kant’s elaboration on what it is to be aware of a subject of thought. In such an awareness, we are aware of the “existence” of a thinker only insofar as that existence is entirely reducible to the necessary organization of particular thoughts and representations. In this way, the deflater incorporates the content Consciousness X into the properties of formal consciousness by deflating the significance of the existence claims under consideration. The awareness of an ‘existence’ described there is just, on this reading, another way to indicate the formal condition on thinking.⁶¹

doesn’t really have a role in Kantian thought. As she reads the vast majority of these passages, Consciousness X is co-extensive with formal consciousness.

⁵⁸ Rosefeldt (2003); Strawson (1966); Brook (1994); Proops (2021). The explanation for the different ways of presenting Option 1 comes down to different ways that scholars understand the ‘formal’ in formal consciousness. The two most popular ways to interpret it are (a) that the formal structure of thought is a *logical* structure, like the way, e.g., *modus ponens* structures particular kinds of contentful arguments; or (b) that it is a *normative* structure, rules by which human beings properly represent relations between concepts and use to reason.

⁵⁹ Kraus, as with almost any reader who has offered a serious analysis of the concept and function of consciousness in Kant (barring Longuenesse, who takes it that practical passages evidence Kant contradicting himself (2017, 152), and Choi (2019)), focuses exclusively on the theoretical passages, and might thereby hold that a different interpretation is required for the practical passages I’ve gathered as instances of Consciousness X. Part of my argument is to show that these passages, both theoretical and practical, invoke almost identical language, terms, and, ultimately, argumentative support, providing good reason to think that whatever interpretation we give of that consciousness in its theoretical context, it should be such that it can be extended to include the practical instances.

⁶⁰ Kraus 2020, 127.

⁶¹ One might object to my use of Kraus as an exemplar for this interpretation since she offers a taxonomy of (what she labels as) five different types of self-consciousness. The first response to this is that, even if she did hold that these five were distinct kinds of self-consciousness, there is substantial evidence to think that she includes the references to Consciousness X in the type of consciousness that tracks what I have been calling formal consciousness. Beyond this, though, it is unclear that she is justified in the taxonomy she offers. The first kind of so-called ‘self’-consciousness she

§2.3. Option 2: Consciousness X is Inflated Formal Consciousness

That is the first way to read Consciousness X as formal consciousness. The second option proffers the same thesis but takes it to have very different consequences. I call this ‘Inflated Formal Consciousness’ and refer to its advocates as inflaters. The inflater embraces the passages that point to awareness of the thing in itself and metaphysically real objects. In order to accommodate them alongside the claim that Consciousness X is an instantiation of formal consciousness, this interpreter *inflates* what it means to be the subject of judgment and thought such that it can include these references. In other words, while the deflater holds that we need to adjust how we read Kant’s substantive metaphysical claims, the inflater instead adjusts the definition of formal consciousness by reinterpreting what it means to be a bare subject of consciousness such that this definition includes those references to the metaphysically real.⁶²

A commitment to inflated formal consciousness is found in Julian Wuerth (2014) and Colin McLear (2021). Both take it that formal consciousness contains or implies the consciousness of one’s noumenal self, viz. the self as a thing in itself or something metaphysically real. Both argue that what it is to be aware of the necessary structure of thought *has* to include the awareness of a real self—in particular, it has to include an awareness of a self or subject that is characterized by

distinguishes is labeled as “transcendental self-consciousness”, but described as having *no* relation or connection to an individual. While I’m on board with the idea that we can describe transcendental consciousness in general terms as a condition on any representation (as Kraus does), insofar as we are considering it in this way we are not referring to an instance of *self*-consciousness—as Kraus herself seems to note, at 108. When we do turn to kinds of *self*-consciousness, it is not ultimately clear what the distinction between what she labels ‘analytic’ and ‘existential’ self-consciousness can amount to when applied to Kant’s text and argument, since they both arise from an act of apperception in a specific empirical thought (see 2020, 127-129). Insofar as there is a difference, it is one that I argue below runs into contradiction.

⁶² As an aside, the fact that both the inflated and deflated readings are advanced, get traction, and can find at least some textual support lends itself to my point that the term and connotation of <transcendental apperception> is much too capacious to allow for interpretive clarity. Part of what generates these two entirely contrary interpretations comes down to what is being read into ‘transcendental’ in the first place.

the unschematized categories of substance, simplicity, standing in certain relations, etc. McLear argues that this must be the case because such an awareness is the only possible explanation of the origin of the *schematized* categories.⁶³ Wuerth holds that this awareness is what justifies Kant in being committed to the substantiality of the soul throughout his mature writings.⁶⁴ This robust awareness of a real self is compatible with Kant's arguments for epistemic humility and his arguments against the rational psychologist in the Paralogisms chapter because, McLear argues, Kant claims only that we are not "acquainted" with the substantial self, where acquaintance should be understood as a technical term indicating recognition of an object in substantial intuition.⁶⁵ On this interpretation, the "emptiness" of formal consciousness is "wholly compatible" with "a kind of subjective content".⁶⁶

In sum, then, the inflater's line on formal consciousness and its relation to Consciousness X is to argue that these references are elaborations on the content of formal consciousness *because* formal consciousness is actually a state with quite significant content. When Kant refers to the content of formal consciousness as purely formal, or "entirely empty", on this view, he means by this only that it is empty of *sensible* content, or sensible intuition. We can, on this view, glean a great deal about the self from these features of thought, at least under a certain mode. The inflater

⁶³ McLear (2020, 2). McLear is motivated to hold this interpretation based on passages like A401-2 and B149, wherein Kant claims that the "objective determination of all representations" (what McLear paraphrases as the categories) is "derived from the principle of the transcendental unity of apperception."

⁶⁴ This commitment, Wuerth argues, is evidenced primarily in Kant's lecture notes (see 2014, 167-8).

⁶⁵ McLear (2020, 27). Wuerth argues for a similar line, focusing on the fact that such an awareness does not amount to *cognition* of the substantial self because we lack a sensible intuition of that self. Nevertheless, according to Wuerth, we might be able to *know*, or come to know that the self or soul has this property (Wuerth 2014, 169-173).

⁶⁶ McLear (2020, 30).

incorporates the content of Consciousness X into the properties of formal consciousness by inflating the notion of ‘empty’ or ‘formal’ to refer to non-sensible content.

§2.4. Option 3: Consciousness X is Empirical Consciousness

The third and final interpretative option for a Two Kinds View is to read Consciousness X as empirical consciousness. This is the least prominent interpretive option of the three, but its advocates can still be identified in the literature.

The clearest proponent of this option is Patricia Kitcher.⁶⁷ She argues that any time Kant is referring to *consciousness* of an *object*, he must have in mind empirical consciousness. For that reason, she takes it Consciousness X—a consciousness that is described as referring to the self as an object—must be empirical consciousness. Since any object of empirical consciousness must be a spatio-temporal appearance, this interpretative option is clearly not intended to generalize over those references to a consciousness of the metaphysically real self, a point Kitcher readily admits.⁶⁸ She argues, though, that we must concede that Kant includes references to the self-consciousness of an empirical existence in explaining the necessary conditions on thinking. Her effort to spell out what kind of representation of existence this could be leads her to argue that the awareness of this existence must be via inner sense—that is, that it is an instance of empirical consciousness.⁶⁹

The way to make this work, Kitcher argues, is to recognize that Kant’s “semiofficial” view is that formal and empirical self-consciousness actually stand in a relation of mutual dependence, where

⁶⁷ Ginsborg 2013 could be taken to be another proponent insofar as she takes formal or transcendental self-consciousness to be conscious of the appearing, empirical, or every-day self.

⁶⁸ Kitcher (2011, 178).

⁶⁹ Kitcher (ibid, 196).

the formal conditions on our thinking require a recognition of our self in inner sense.⁷⁰ Taking this seriously reveals a third option for interpreting Consciousness X on a Two Kinds View. If we hold that Consciousness X is awareness of the self as an object but deny that we can be in any way aware of non-empirical objects, we arrive at the conclusion that Consciousness X is an instance of empirical consciousness. The implication of this is that our awareness of the formal conditions of thinking are not *a priori*, but rather derivative on—or, at minimum, the awareness mutually depends on—awareness of our self in inner sense.

I have canvassed three different options for reconciling the claim that Kant only has two kinds of consciousness (the Two Kinds View) with the passages that seem to refer to a consciousness (Consciousness X) with features that are in tension with the properties of the respective kinds. Option 1: Consciousness X is formal consciousness, where the purported awareness of a further existence is understood as an elaboration on what it means to be aware of a bare subject of thought (the Deflated View). Option 2: Consciousness X is formal consciousness, where awareness of an “empty” or bare subject is taken to mean empty of *sensible* content, and in fact includes fairly significant non-sensible content (the Inflated View). And Option 3: Consciousness X is empirical consciousness: the awareness of our existence is of our existence as spatiotemporal agents.

Clearly, these options pull in very different directions, and both textual and philosophical support can be found for each. Luckily, we do not need to settle between them. All three, I argue, run into serious problems when we attempt to actualize their commitments in key arguments in Kant's account of self-consciousness. Given this, I conclude that we should abandon the Two Kinds View

⁷⁰ Kitcher (ibid, 160).

by committing to a fourth interpretive option: that Consciousness X is a distinct, third kind of consciousness. I now turn to that argument.

§3. The Need for a Third Kind of Consciousness in the Transcendental Deduction

Kant is clearly committed to two kinds of self-consciousness: formal self-consciousness, and empirical self-consciousness. The question is: do these two kinds exhaust his theory? We have now seen three different ways to answer this question in the affirmative. In light of a series of passages that introduce a kind of self-consciousness that does not clearly belong to either one of the two previously identified kinds, we can (1) assert that they are references to formal consciousness, where that means deflating Kant's references to a non-sensible existence; (2) assert that they are references to formal consciousness, where that means inflating the 'empty' content of that state; or (3) assert that they are references to empirical consciousness.

I will argue here that none of these three routes is tenable: an attempt to integrate the properties of what I have called Consciousness X into any of these interpretations results in a state with incompatible or contradictory properties. I will do so by showing that we cannot make sense of a key—and notoriously difficult to understand—argument at the heart of one of Kant's main discussions of self-consciousness, the Transcendental Deduction, if that argument appeals to deflated formal consciousness, inflated formal consciousness, or empirical consciousness. The argument, in other words, must appeal to a third, distinct kind of self-consciousness.

I devote this entire section to this argument from the Deduction: it is, as it were, my star witness for the Three Kinds View. There are three reasons for this. First, the section of the text I focus on here, §25, is a notorious one for any account of consciousness and self-consciousness in Kant. It is

frequently brought up as evidence against some view or another, whatever the content of that view is.⁷¹ Perhaps for that reason, many interpreters ignore it entirely, and few have offered an in-depth account of the specific argument of this section. In looking at this argument first, I show what pieces we need to understand it and then go on to argue that those pieces are found in Kant's account. Second, this argument comes in the heart of the chapter wherein Kant purports to offer his positive theory of self-consciousness: it stands to reason that it offers important information for how to understand that theory. Third, this is a key *theoretical* argument for consciousness and self-consciousness. I have been arguing that Kant's theory of self-consciousness is continuous—at least in form—between both theoretical and practical works. Traditionally, though, readers have taken the practical invocations of consciousness to float free of Kant's 'serious' or 'proper' theory as it is spelled out in the theoretical texts. To properly motivate my view that there are three kinds of consciousness throughout his theory, and therefore to motivate that Kant consistently uses the term 'consciousness' with intention, it is important to show that that view is articulated in the *Critique of Pure Reason* itself.

I here articulate an argument form that invokes a third kind of consciousness, an argument that I contend we can see Kant utilizing or relying on throughout the first *Critique*. This section goes into the detail of this argument and its consequences. §4 will provide further instances where Kant appeals to this same form of argument, further evidencing my claim that a third kind of consciousness is essential to his account in theoretical as well as practical contexts.

⁷¹ Notably, the only place in Longuenesse (2017) where she articulates the view that Kant has three kinds of consciousness or self-consciousness is in considering §25 of the Deduction. It also informs Kitcher's (2021) criticism of Kraus' account, and Kraus considers it at length as a potential source of objections to her account (2020, 124-7).

§3.1. The No Real Self-Cognition Argument

The Deduction is Kant's attempt to explain why *a priori* concepts can have necessary application to sensible objects. The main move in this explanation is to account for the conditions of human thinking. And a central part of *that* account is making sense of the ways that a human can be conscious of herself. That includes explaining the conditions on self-cognition. In essence, Kant says that, just as our cognition of outer objects is circumscribed by space and time as the forms of our intuition, so too is our cognition of our self. "Through inner sense", he writes,

we intuit ourselves only as internally affected by our selves, i.e., as far as inner intuition is concerned we cognize our own subject only as appearance but not in accordance with what it is in itself. (B156)

So far, so in keeping with the epistemically humble Kant: we can't cognize anything as it is in itself, not even our own self. But in this context, it is all the more striking that Kant immediately turns (§25) to discuss a contrast case. I'll call the argument of this section **No Real Self-Cognition (NRC)**, and title its parts for clarity. Kant begins by identifying a consciousness that we *do* have.

Necessary Conditions on Real Self-Cognition:

[...] in the synthetic original unity of apperception, I am conscious of myself not as I appear to myself, nor as I am in myself, but only that I am. [...] Now since for the cognition of ourselves, in addition to the action of thinking that brings the manifold of every possible intuition to the unity of apperception, a determinate sort of intuition, through which this manifold is given, is also required, my own existence is not indeed appearance (let alone mere illusion), but the determination of my existence can only occur in correspondence with the form of inner sense, [...] and I therefore have no cognition of myself as I am, but only as I appear to myself. The consciousness of oneself is therefore far from being a cognition of oneself, [...]. (B157-8)

Kant is here arguing that, even though we *do* have a consciousness "that I am", we do *not* have cognition of our real self. To show this, he isolates two jointly necessary conditions on self-cognition. In order to cognize yourself, you must have both (i) the "action of thinking", parsed both above and below as a *consciousness* of your self, and (ii) a determinate intuition of your self.

Human minds fail to meet (ii), and so human minds fail to self-cognize. Or, importantly, human minds fail to cognize their *real* self, because they lack an intuition of their *real* self. As we learned immediately ahead of this passage, humans do intuit their selves as they appear, and so do cognize their appearing self.⁷²

I want to here open a brief sidebar into the nature of Kantian cognition [*Erkenntnis*], as a reader might be surprised by my identifying *consciousness* and *intuition* as the two elements of a cognition. I take it that these conditions are jointly necessary for (human) self-cognition, but they are not *sufficient*. Famously, Kant argues that cognition depends on both a sensible intuition and the application of *concepts* (see A51/B75).⁷³ I do not mean to be here overriding this account nor undermining any additional further conditions that are integral for Kantian cognition.⁷⁴ Likewise, though, I am not suggesting that consciousness here *plays the role* of the concept condition.⁷⁵ This

⁷² How exactly we are supposed to cognize our self or subject as it is appears and is affected is a hard question that I am bracketing here. It is a matter of a great deal of scholarly debate and deviation: questions include whether inner sense is directly analogous to outer sense, and how exactly we perceive ourselves in inner sense. For recent discussions that explore this, see Sethi 2021, Kraus 2020, Chignell 2017, Valaris 2008, and Emundts 2006.

⁷³ I have a cognition of my pen because there is an existing object that is present or given to my mind, that I am capable of describing with general features (the pen is blue, it is thin, it is light, etc.). Intuition is the contribution of the mental faculty of *sensibility*, while concepts and thinking are the contribution of the mental faculty of *understanding*. In virtue of being a result of the relation between intuition and concepts, then, cognition is a result of the relation between two distinct mental faculties. This is all summarized by Kant with a famous dictum: "Thoughts without content are empty, intuitions without concepts are blind".

⁷⁴ Interpreters generally agree that in addition, cognition demands a 'real possibility condition' and a 'determinate content condition'. See Chignell (2010), Watkins and Willaschek (2017), Hogan (2009), Kain (2010), Smit (2009), and Schafer (2022) for different ways of spelling out either or both of these conditions.

⁷⁵ There is a way in which we can read Kant as taking the concept condition on cognition to be interchangeable with a consciousness condition. Look at his claim, for instance, in the *Jäsche Logic* that "All our cognition has a *twofold* relation, *first* a relation to an object, *second* a relation to a subject. In the former respect it is related to *representation*, in the latter to *consciousness*, the universal condition on all cognition in general." (JL 33). I have argued that Kant is committed to a non-referring basic kind of consciousness: formal consciousness. It is non-referring because it functions as the necessary and formal condition of all thinking and representing in general. Insofar as there is a consciousness, viz., formal consciousness, that is the condition on all thinking and representation, and insofar as cognition is a species of thinking or representing, then there is a trivial (or perhaps more appropriately, transcendental) sense in which consciousness functions as a condition. Formal consciousness is the general condition on thinking: ipso facto it is a general condition on cognition. This is the 'consciousness condition' that I take Kant to be articulating in the *Jäsche Logic*. It is worth noting that a kind of consciousness plays such a role in Kant's account of thinking, but this is not a role that I take to be unique

argument, on my view, is part of an important and (as far as I am aware) unrecognized element of Kant's account of cognition, viz. that a referring consciousness is an *independent* necessary condition on cognition, in addition to the intuition condition, and in addition to the concept condition.⁷⁶ I cannot defend this interpretation at length here. I take it that the structure of the argument I uncover provides enough evidence to motivate my point for present purposes. But I want to gesture at this account, and the fact that taking seriously the argument in §25 reveals something both about the kinds and nature of consciousness that Kant is committed to, and about the structure and constitutive elements of cognition.⁷⁷

Coming back to the main discussion now. The argument looks like this. There are two necessary conditions on real self-cognition: consciousness of the self, and an intuition of the self. Human minds *do have* the former, but they do not and cannot have the latter, and so they do not and cannot cognize their real self. For our current discussion, the most important feature of this argument is that its structure relies on all the mental states in question—consciousness, intuition, and cognition—having the same referent. That Kant relies on appeal to one consistent referent is

to his account of *cognition*. Insofar as systematic accounts of Kantian cognition have recognized a distinctive role for cognition, they have focused on this general consciousness constraint. See Tolley (2017, 3321) and Smit (2000, 265). This is not the kind of consciousness condition I am interested in spelling out here: I take it that Kant is here appealing to a more specific and specifically referring consciousness condition. See n.78 for further discussion.

⁷⁶ And in addition to the real possibility condition, and in addition to the determinate content condition. That a referring consciousness is a constitutive element of Kantian cognition is recognized by almost all accounts, it is just folded into the 'givenness' or intuition condition (see, e.g., Watkins and Willaschek (2017, 89-90) and Schafer (2022, 255). This argument, I take it, provides good reason to pull it apart.

⁷⁷ I will note that in almost all theoretical cases, the givenness of an object and consciousness of it go hand in hand, and so this disambiguation is in some sense just (interesting!) bookkeeping that helps to specify the special epistemic status of consciousness and cognition both. But understanding a referring consciousness to be a distinct element of cognition is instrumental, I contend, in recognizing why *practical cognition* is still a form of *cognition*, and why it can be of supersensible objects. I discuss this in more detail in the dissertation's conclusion. For now, I leave this open as a future direction for research.

supported by two further reiterations of the argument, one by way of an analogous explanation to object cognition and the other by way of a counterfactual case (titles are my addition).

Analogy to Object Cognition:

Just as for the cognition of an object distinct from me I also need an intuition in addition to the thinking of an object in general (in the category), through which I determine that general concept, so for the cognition of myself I also need in addition to the consciousness, or in addition to that which I think myself, an intuition of the manifold in me, through which I determine this thought; and I exist as an intelligence that is merely conscious of its faculty for combination but which, in regard to the manifold that it is to combine, is subject to a limiting condition that it calls inner sense [...]. (B158-9)

Counterfactual Case:

[...] and that [my mind] can therefore still cognize itself merely as it appears to itself with regard to our intuition (which is not intellectual and capable of being given through the understanding itself), not as it would cognize itself if its intuition were intellectual. (B159)

In **Analogy to Object Cognition**, Kant articulates that cognition of the real self functions in the same way as cognition of external objects. For the cognition of an external object *x*, I require an intuition of the object of cognition *x*, and a concept of that very same object *x*. Likewise, cognition of the real self depends on intuition of that self, and a consciousness of that same self.⁷⁸ In **Counterfactual Case**, Kant asserts that *were we to have a different intuition* (were it the case that human intuition was intellectual and not discursive), we *would* cognize our real self. Following the line of thought presented in **Necessary Conditions** and **Analogy**, this is only a possible counterfactual if a consciousness (the other condition on self-cognition) is already in place. There are two conditions on cognition: if meeting the second condition (having an intuition of the object)

⁷⁸ This analogy may seem to rely on a parallel between (and thereby an interchangeability of) 'concept' and 'consciousness', contrary to what I denied in n.75. Further reason to think that the concept and consciousness conditions are *not* interchangeable in this way can be found in Kant's concept of "obscure representations". Kant describes these (JL 33) as representations *without* consciousness, and his primary example is the Milky Way. When you see the constellation in the sky at night, it appears to you as a single mass, even though you are in fact perceiving the multitude of stars that compose it. You clearly, though, have the *concepts* of the stars that compose it. For obscure representations to be cogent, the concept and the consciousness conditions cannot be coextensive.

would mean that an agent would attain cognition, then it must be the case that this agent already meets the first condition.

Kant is clearly committed to the following line of thought. Cognition of the real self is possible iff an agent has an intuition of that self *and* they are conscious of that self. Human minds *are* conscious of that self, but they cannot intuit that self, and so real self-cognition is impossible. This line of thought is only cogent if the consciousness, the intuition, and the cognition under discussion all refer to the same object.

Now, in the context of the broader argument I began in chapter 1, one thing that leaps out is that this entails Kant's commitment to an agent being conscious of her real self. Since the referent of the cognition in question is the real self, and since the argument depends on all of the mental states it invokes having the same referent, it should follow that the consciousness "only that I am" is consciousness of the real self.⁷⁹

In keeping with my argumentative strategy here, though, we can say that this consciousness—the consciousness that I am—shares the features of what I have been describing as Consciousness X. I've here identified an argument in Kant that depends on a consciousness that has the following feature: that it refers to an object (a self) that we *could* but *do not* have a corresponding intuition of. To secure the conceptual distinctness of the three kinds of consciousness, I want to focus on the role of this consciousness in the structure of this argument. If the consciousness in question

⁷⁹ Wuerth likewise recognizes the peculiar way in which Kant contrasts the consciousness "only that I am" from the consciousness "as I appear in myself" and "as I am in myself", and likewise comes away with the conclusion that this is a consciousness of our real self (2014, 122). What I am providing in this section is an argument to substantiate that move, since, absent such an argument, an opponent of the view will just insist that you are assuming what they deny.

does not have this feature, then the conclusion Kant draws about real self-cognition does not follow from the relationship between consciousness and intuition as it is supposed to.

Can the content of the consciousness identified here be explained as part of the content or function of either formal consciousness (as it is understood either by the deflater, or by the inflater), or of empirical consciousness, while still playing the same role in the argument? If it can be, then this is good reason to think that it is not conceptually distinct. I will now argue that it cannot be explained by any of these interpretations.

§3.2. Why NRC Cannot be Appealing to Deflated Formal Consciousness

First, what I have called deflated formal consciousness. Recall that I originally defined formal consciousness as a mental state with *no* representational content: it is a structural or formal feature of thought only. This definition entails that the consciousness is non-referring: it does not have an object or content. The deflater, in interpreting formal consciousness, leans into this feature. Any instances where Kant seems to be suggesting that there is a substantive or concrete awareness of our existence they read as elaborating on what it means for an act of apperception or mental activity to provide the formal conditions on one's collection of representations: spelling out what it means to have this kind of structure in thinking.

Plugging this account of consciousness in for the Consciousness X at issue in the No Real Self-Cognition (NRC) argument yields the following picture. We are conscious of the formal features of our thinking. That is, we are conscious that our minds necessarily function in certain ways as to allow us to have discrete thoughts and representations, and to recognize those thoughts as

having a common subject. We do not have an intuition of a metaphysically real self, and so we cannot cognize such a metaphysically real self.

On this picture, the consciousness at issue is entirely disconnected from the object of intuition and cognition. Given this disconnect, the way Kant presents the argument fails to make sense. It renders wholly mysterious the references to “the existing intelligence” that is the object of consciousness. And it undermines the counterfactual claim. Were we to have an intuition that corresponds to the object of our consciousness, then we would have cognition of the real self. Since formal consciousness does not refer to an object at all, we could not—even hypothetically or counterfactually—have an intuition of its object.⁸⁰

If, in light of this argument, we say that formal consciousness *does* have an object, then the minimal definition of formal consciousness is threatened. The properties of the state would be contradictory: it would both have and not have a referent. If deflated formal consciousness is the relevant kind of consciousness in §25, then the argument Kant makes there cannot be made coherent without sacrificing the internal coherence of this kind of consciousness.⁸¹ Given this, we

⁸⁰ Kraus argues explicitly that formal consciousness does not become cognition because this consciousness does not refer to *any* content or contentful state, *not* “from the fact that the ‘I think’ yields a content which is, as it were, hidden from view” (Kraus and Freitag 2022, 114). But a description of a content hidden from view seems to be precisely what Kant is appealing to in §25. In her book, Kraus responds to this kind of objection by arguing that when the ‘I think’ is “actualized in an act of apperception in a mental episode leading to an occurrence in empirical consciousness”, then it expresses “the fact that this form is actualized by an existing thinker” (Kraus 2020, 125). If it is the case that this is the way to explain these existence claims, then the existence referred to in §25 is a particular empirical existence (namely, mine), and so the consciousness under discussion has to empirical consciousness. In that case, the criticisms I raise in 3.4. apply to this version of Kraus’s view.

⁸¹ Beyond the specific problems each run into, if we were to take it that Kant is relying on either kind of consciousness then he has formulated those claims in the most confusing way possible. If we have no epistemic relation whatsoever to our real self, denying that we have the particular epistemic relation of cognition to that self should follow from that alone. If either kind of consciousness is what Kant had in mind here, one would hope he would appeal to this straightforward point. This point is even more pressing if one were to argue that there is no real self at all—that is, that ‘things in themselves’, for Kant, are only epistemic, limiting conditions on our knowledge. In that case this reads as if Kant is trying to kill a fly with a bazooka when a newspaper is sitting right there. If there is no real self, then there is

should rule this out: formal consciousness, at least on this deflated reading, cannot be the kind of consciousness at stake here.

§3.3. Why NRC Cannot be Appealing to Inflated Formal Consciousness

The explanation of the consciousness in NRC cannot be deflated formal consciousness because Kant's argumentative strategy is not comprehensible if the state in question inherently lacks a referring object. In light of this, a natural move to make is to hold that this consciousness *does* have an object to which it refers. If we want to read Consciousness X as formal consciousness, then this leads naturally to the second interpretative option: inflated formal consciousness. The inflater takes formal consciousness to be 'empty' only of sensible content. On this reading, the state *does* refer to a non-sensible object or subject. Importantly, the inflater's reading relies on this referring state being coextensive with the formal structure of thought: this is how our awareness of our non-sensible existence generates the categories and justifies their application.

The first problem with inflated formal consciousness in the context of §25 of the Deduction lies in the very first step of the account. Kant claims that I am conscious *only that* I am and denies that this is the same as either consciousness of myself as I appear, *or* consciousness of how I am in myself. If this consciousness is supposed to be an awareness of my noumenal self as instantiating certain properties (like substantiality) as unschematized categories, then Kant ought to have said

certainly no real self-cognition. Minimal charity suggests that we should take seriously that the two pages describing and reiterating the conditions of real self-cognition are in service of detailing the more complicated state of affairs wherein we are conscious of, but do not cognize, our real self.

no such thing. On this interpretation, this claim is rendered at best extremely puzzling, and at worst either misguided or deliberately obfuscating.⁸²

But an attempt to walk back the significance of the content contained in inflated formal consciousness reveals a deeper problem with identifying formal consciousness with a referring state. Several of Kant's main arguments for the nature and function of self-consciousness—most notably, his arguments against the rational psychologist in the Paralogisms—rely on the thesis that no further information can be gleaned by simply reflecting on the place and function of the 'I' (either as an occurrent expression, or in simply reflecting on the need for a subject place in any thought). If we were to hold that formal consciousness does have this inflated definition, even minimally (that we have an only indeterminate representation of the real self), then the rational psychologist does not seem to be in the grip of a fallacy at all.⁸³ Yet again, then, while accepting that Kant's argument in §25 relies on (a certain kind of) inflated formal consciousness might be enough to make the argument as presented work, it does so at the expense of several of his other arguments for the nature of self-awareness. We do not want to explain the nature of this kind of consciousness *at the expense of* the importantly non-referring structure of thought.

⁸² A similar criticism is raised by Proops (2021, 115). As Proops notes, as well, Kant is not here denying that we have cognition or knowledge of how we are in ourselves: he is denying that we are *conscious* of it. This forecloses an inflator's reply that we can make substantive claims about the real self as long as we are clear that this access is not under the mode of cognition.

⁸³ A full account of the argument structure in the Paralogisms, and in particular where Kant locates the fallacy he attributes to the rational psychologist, is beyond my scope here. What I will say here is that in critiquing the rational psychologist, Kant relies on the fact that, from reflection on thought alone, we cannot rule out whether we are (for example) a substance or an accident. Even if McLear and Wuerth are right that Kant never allows that we can cognize the content expressed by the positive proposition, it must follow on either of their views that we can cognize the content expressed by the negative proposition (viz. I am *not* an accident).

If inflated formal consciousness is the relevant kind of consciousness in §25, then it is both the case that Kant begins the argument with a mistake or obfuscation, and he perpetuates this kind of mistake or deliberate misleading of the reader in pointing out the fallacies of his opponents. We should likewise reject this interpretation: formal consciousness, on this inflated reading, cannot be the kind of consciousness at stake here. And since the previous section showed that we cannot insert *deflated* formal consciousness into this argument, we can conclude that the argument does not appeal to formal consciousness.

§3.4. Why NRC Cannot be Appealing to Empirical Consciousness

The only option that remains to maintain a Two Kinds View of consciousness is to read Consciousness X in §25 as empirical consciousness. But this reading also fails to make sense of the argument. Empirical consciousness is consciousness of spatiotemporal representations, and so empirical self-consciousness is consciousness of the self as it appears or as an object of experience. But this is, at least by Kant's assertion in the immediately preceding section, a self that we *can* cognize. Yet again, we are in a position where this interpretation fails to make sense of the feature of consciousness that is necessary for the argument, i.e., that it has the same referent as the potential intuition and cognition under discussion. Further, applying this interpretation flies in the face of Kant's explicit assertion that this consciousness is consciousness "not of myself as I appear to myself". Taken altogether, if we were to read Consciousness X here as empirical consciousness, it would likewise result in a state with contradictory properties:

empirical consciousness would both be awareness of the self as an experienced object such that it could be cognized, *and* awareness of a self with properties such that it could never be cognized.⁸⁴

The consciousness at issue in No Real Self-Cognition cannot be formal consciousness, and it cannot be empirical consciousness. We do need these two kinds of consciousness in our taxonomy. But if we are *limited* to these two kinds of consciousness, then we do not have a satisfactory reading of the argument for our inability to cognize our real self. An attempt to read that consciousness as either previously identified kind results, on the one hand, in an incoherence in Kant's argument, and on the other in describing a mental state with contrary properties. Since the argument that generates these problems comes at such a central place in the Deduction's discussion of consciousness, it is unlikely that Kant was confused or mistaken about his description of this consciousness. The interpretive option that remains to us, I argue, is to reject the contention of the Two Kinds View and hold instead that there is a third kind of self-consciousness in Kant's account of the mind.

§3.5. Takeaways and the Need for Three Kinds of Consciousness

A commitment to the Two Kinds View results in confusion and contradiction in the heart of the Transcendental Deduction. But there *is* something importantly right about each of the interpretations that attempt to preserve it. As the proponent of empirical self-consciousness notes, Kant is clear when he is describing a consciousness with representational content. The inflator is

⁸⁴ Further, if one were to insist, as Kitcher suggests we must, that any reference to object self-consciousness has to be a reference to empirical self-consciousness, then both the argument for No Real Self-Cognition and Kant's arguments more generally for the transcendental conditions on thinking run into a vicious circularity or regress problem. Kitcher (2011, 197) suggests that the remarks in §25 are Kant trying to work out the priority relation between empirical and formal consciousness, and concludes that a representation of my self as an object of experience actually does precede my awareness of myself as a subject of thought. But this gets Kant's core claim (e.g. B133) that an analytic unity of apperception *always* presupposes a synthetic unity backwards. For more on this objection, see Kraus (2020, 97-8).

right to point out that Kant indicates we have an awareness of our non-sensible self: that there is a consciousness with supersensible content. And the deflater is right to insist that a non-referring consciousness is essential to understanding Kant's positive arguments for the structure of thought and negative arguments critiquing the rationalist.

Together, these insights and careful attention to the argument in §25 of the Transcendental Deduction reveal an appeal to a distinct kind of consciousness. I've argued that we can generalize from the distinction between formal and empirical consciousness to an individuating criterion: we can identify a distinct kind of consciousness if its primary properties are incompatible with the primary properties of a previously identified kind of consciousness. We now have the material to apply this criterion again, with the result being that Kant is committed to *three*, and not two, kinds of consciousness. Kant's account of consciousness is best encapsulated by this Three Kinds View. I now turn to gather further evidence that Kant relies on the distinction between three kinds of consciousness in the *Critique of Pure Reason*.

§4. Further Evidence for Three Kinds of Consciousness

We have now seen an argument that relies on a distinct, third kind of consciousness in Kant's account of the mind. This argument is particularly compelling, I take it, because it is situated in such a central part of Kant's theory of self-consciousness. Nevertheless, I do not expect the No Real Self-Cognition argument to be entirely independently satisfying. If Kant is committed to the Three Kinds View, it should be the case that this view—both the argument structures that rely on it, and this third kind of consciousness—is appealed to throughout the *Critique of Pure Reason*. In this final section, I argue that this is the case. I show that three further passages in the first *Critique*

that are taken to be opaque, inchoate, or inconsistent are made clear and understandable once we employ the resources of the Three Kinds View. Extending the insights of the view in this way will in turn offer further evidence for my claim Kant is appealing to three distinct kinds of consciousness.

§4.1. No Real Self-Cognition in the General Remark in the B Paralogism

The first of these passages comes from the final section of the revised, 'B' Paralogism chapter. This section, titled 'General remark concerning the transition from rational psychology to cosmology' has long puzzled commentators.⁸⁵ It is taken that the section vacillates between extremely minimal claims about the function of the 'I' and substantive metaphysical and moral pronouncements without explicating the connection between the two.

On my view, this puzzle is resolved by recognizing that Kant is again here reiterating his argument against real self-cognition. Once you know what you are looking for, it is evident that the argument structure is almost entirely parallel. Recall that the key features of that argument were [i] Necessary Conditions of Real Self-Cognition, [ii] Analogy to Object Cognition, [iii] Counterfactual Case. The following passage, which represents the bulk of the discussion in the General Remark, contains all three. The text numbering is my addition, to mark when the presence of these features.

But the proposition "I think," insofar as it says only that I exist thinking, is not merely a logical function, but rather determines the subject (which is then at the same time an object) in regard to existence, and this cannot take place without inner sense, [...]. It is in this latter that the thinking self must now seek the conditions of the use of its logical functions for categories of substance, cause, etc., so as not merely to indicate itself as object

⁸⁵ Compare, for example, Longuenesse, who argues it is a sharp deviation from the rest of the chapter to gesture at Kant's practical views (152-3) with Kraus, who argues that it is Kant's development of the empirical character of the "I think" and our cognition of the self in inner sense (2020, 127; 131; 153).

in itself through the “I,” but also to determine its kind of existence, i.e., to cognize it as a noumenon; [ii] which, however, is impossible, since inner empirical intuition is sensible, and makes available nothing but data of appearance, which [i] affords nothing for knowledge of the separate existence of the object of pure consciousness, but can serve merely in behalf of experience. But [iii] suppose there subsequently turned up—not in experience but in certain (not merely logical rules but) laws holding firm *a priori* and concerning our existence—the occasion for presupposing ourselves to be legislative fully *a priori* in regard to our own existence, and as self-determining in this existence; then this would disclose a spontaneity through which our actuality is determinable without the need of conditions of empirical intuition; and here we would become aware that in the consciousness of our existence something is contained *a priori* that can serve to determine our existence, which is thoroughly determinable only sensibly, in regard to a certain inner faculty in relation to an intelligible world (obviously one only thought of). (B430-1)

Just like in §25, Kant reiterates [i] that to have cognition of the real or noumenal self, one would have to have *both* a consciousness of that self, and an intuition of that self. But since the only available intuition of a self that we can have is a sensible intuition in inner sense, we *cannot* have a cognition of this self. This is despite the fact that we *do* have a consciousness of the “separate existence” of this self, understood as real object. That we do have this consciousness is substantiated, yet again, by a hypothetical or counterfactual claim [iii]: “suppose there subsequently turned up” “the occasion for presupposing ourselves to be legislative in regard to our own existence”. If this were to be the case, then we *would* be able to determine the existence that we are, for the sake of the present discussion, only conscious of.⁸⁶

It is worth noting that the counterfactual claim is here presented with an interestingly different addition, namely that it suggests that we *could* in fact come (and in fact, when taken together with Kant's practical arguments, we *do* come) to recognize these laws, and therefore could in fact come

⁸⁶ This language of when and how our existence is determined or determinable echoes the lengthy footnote in §25: e.g., “Now I do not have yet another self-intuition, which would give the determining in me, of the spontaneity of which I am only conscious, even before the act of determination [...] thus I cannot determine my existence as a self-active being.” (B157-8n).

to have at least a kind of cognition of our real self. Understanding this difference, I take it, goes hand-in-hand with identifying the specific role consciousness plays in cognition (as noted in §3). I discuss possible reasons for this difference and the role of consciousness in theoretical and practical cognition in the dissertation's conclusion. For now, I merely note the difference and that it does not change the structure of the claim, viz. that, if other conditions were to obtain, we could cognize the content of the consciousness that we *already have*.⁸⁷

In essence, then, the General Remark presents the same argument for No Real Self-Cognition (NRC). Since, as we have seen, that argument relies on a third distinct kind of consciousness, we can conclude that the very same consciousness is being appealed to here. Since the Paralogisms is the other chapter of the first *Critique* where Kant discusses consciousness at length, this is as you would expect if he were committed to the Three Kinds View. The presence of the NRC argument in the Paralogisms provides further proof of his commitment to a tripartite theory of consciousness.

§4.2. Three Kinds of Consciousness at Work in the Critique of Pure Reason

This first piece of evidence showed that Kant returns to an argument for which a third kind of consciousness is essential, viz. the No Real Self-Cognition argument. This substantiates both that this argument plays an important role for him, and the presence of this third kind of

⁸⁷ Kant immediately goes on to say that such a state of affairs would “nonetheless bring all the attempts of rational psychology not the least bit further” (B431). This is more evidence that the inflator's picture of formal consciousness cannot be right, because if we came to recognize the real self via the very same mechanisms and consciousness that produced the expression ‘I think’, then this *would* substantiate the attempts of rational psychology, albeit via a somewhat different route.

consciousness in his thought. Here, I will give two examples of the work a Three Kinds View can do by presenting two claims of Kant's that are clarified when we assume the tripartite account.

The first of these comes from the Resolution to the Third Antinomy, where Kant is explaining how the actions of human beings can be both naturally determined (there is an explanatory or generative cause for every effect) and free. This is possible, Kant argues, only on the assumption that some actions have their source outside of the natural, deterministic world. In a striking passage, he claims that we are *aware* of the source of these undetermined actions.

In the case of lifeless nature and nature being merely animal life, we find no ground for thinking of any faculty which is other than sensibly conditioned. Yet the human being, who is otherwise acquainted with the whole of nature solely through sense, knows himself also through mere apperception, and indeed in actions and inner determinations which cannot be accounted at all among impressions of the senses; he obviously is in 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. (A546-7/B574-5)

Kant asserts both that there is a part of the human being that is outside of nature (in particular, certain mental faculties are merely intelligible objects), and that we are aware that they are such—viz., that they are capable of actions that cannot be deterministically explained. With the resources of the Three Kinds View, we can read these remarks as continuous with the theory of self-consciousness that Kant has been espousing throughout the *Critique*.⁸⁸

The second example is a well-known passage from early on in the B edition of the Transcendental Deduction. Kant is here spelling out what it means to have a unified set of representations and why the unity of thought is a necessary precondition on representations.

⁸⁸ As opposed to either here introducing (unprompted) a radically new idea or access that we have to ourselves, or slipping up and reverting to a previous commitment to a more rationalist metaphysical system (as is argued by, e.g., Ameriks (2000, 208-213).

The thought that these representations given in intuition all together belong to me means, accordingly, the same as that I unite them in a self-consciousness, or at least can unite them there in, and although it is itself not yet the consciousness of the synthesis of the representations, it still presupposes the possibility of the latter, i.e., only because I can comprehend their manifold in a consciousness do I call them all together my representations, for otherwise I would have as multicoloured, diverse a self as I have representations of which I am conscious. (B134)

What is notable here is the assertion that, absent a unified manifold for consciousness, I would have as *diverse* a self as I have representations. As the latter claim about the unity of representations is commonly read—in particular, how it is read by the proponents of Two Kinds Views—Kant should argue that absent such a unified manifold I have *no self at all*: the structural unity of my representations is what gives rise to a self in the first place. We require an explanation, then, for why he instead endorses a claim about an untenable multiplicity of selves.⁸⁹

The Three Kinds View is well-suited to offer such an explanation. On the view I've put forward, imagining that we do not have an awareness of our self as the formal subject of thinking, we would still have an indeterminate awareness that we have a self. But since we would not have any formal or logical identity amongst our representations, it makes sense that we would have a kind of target-confusion when that indeterminate awareness is combined with particular sensible representations, where that confusion would manifest as the *experience* that we have a self for each particular representation. This is not, of course, a situation in which we ever do or could find ourselves in, which is part of the reason Kant asserts that the formal structure of thought is a

⁸⁹ I say 'untenable' because it is clearly in the background here that having as many selves as one has representations would obliterate the concept of what it is to be a self at all: there is some criteria of continuation or identity that is baked into the concept. I do argue that Kant endorses a metaphysics of the self on which there are multiple selves, and so I take it that there is a line between a self as diverse as my representations, and a commitment to exactly one self. I discuss this in detail in chapter 3.

transcendental condition on thinking in general. But it is only with a third kind of consciousness in hand that we can make sense of the case where this transcendental condition does not obtain.

I've argued that three kinds of consciousness are appealed to in Kant's core arguments for self-consciousness. I've here shown that recognizing a commitment to these three kinds allows us to explain passages that are otherwise unclear or appear inconsistent. There is good textual support for a Three Kinds View of consciousness, and a commitment to this view has both interpretive and philosophical payoffs.

§5. Conclusion

I have argued that careful attention to Kant's arguments for the different ways that we understand and represent ourselves reveal that he relies on three, conceptually distinct kinds of self-consciousness. While the vast majority of readers have read Kant as committed to a Two Kinds View of consciousness, I've shown that he is better understood as having a Three Kinds View.

In the course of this argument, I offered definitions of two of these three kinds: formal self-consciousness, understood as awareness of the non-referring structure of thought, and empirical self-consciousness, understood as awareness of your self as an object of experience. To fully substantiate the view, then, we require a definition of this third kind of consciousness. This is exactly what I offered in chapter 1. This third kind of consciousness is what I have called real consciousness: a consciousness that has an intentional object, but one that is not directly represented. Taking these two arguments together, then, satisfies all the commitments of this mental state that I set out to explain. Understanding real consciousness as a feeling-towards allows us make sense of it as a mental state consistent with Kant's idealist commitments. And

understanding its properties as such reveals that Kant appeals to three distinct kinds of self-consciousness, and that each of these kinds is well-defined.

In understanding this real self-consciousness to be a conceptually distinct, third kind of consciousness present in Kant's account of the mind, I take it that we can isolate a criterion for distinguishing amongst all of Kant's different references to consciousness, self-consciousness, and various cognates, like *apperception*. As I've discussed, Kant readily uses apperception to describe self-consciousness of various kinds but is unclear about the significance and difference between different ways of describing apperception (*pure, transcendental, mere*, etc.). Identifying the content and difference between real, formal, and empirical consciousness provides a tool with which to go back into text and properly identify and define his technical terminology.

Such a tool will likewise function to clarify key arguments that otherwise seem confused or confusing. I've begun that work here in employing the Three Kinds View to clarify §25 of the Deduction, the General Remark to the B Paralogism, Kant's claim that I am aware of my ability to determine actions and events via certain mental faculties, and his assertion that absent a unified manifold of my representations, I would have multiple (instead of no) selves. An important further extension of this work is in showing how Kant's theory of self-consciousness is developed across both theoretical and practical writings. Offering a full accounting of how to understand the three kinds of consciousness, their instantiations, and their relationship, is a complex and lengthy project: I look forward to spelling it out in future work.

Here, I've argued that Kant is committed to the Three Kinds View of consciousness. His account of the mind and of how we represent and understand the self relies in important ways on these three, distinct, kinds of consciousness.

Chapter 3

Are Two Selves Too Many Selves?

Puzzles about the Relation Between Selves in Kant

If I told you that you were two selves, or two people, your reaction would likely be one of confusion. In so stating, I appear to be in the grip of a misunderstanding: what it is to be *you* is to be exactly one person or self. Absent this singularity, we have a hard time grasping what I am referring to. “But which one is actually *me*?” seems both a natural and fair response. Philosophical intuition and commonsense both appear to support the view that we have, at most, one self.

Whether Kant subscribes to this intuitive view is a matter of great debate. In this chapter, I argue that he most likely does not: the best reading of Kant’s metaphysics of the self posits two entities that stand in an intimate relation. This, on my view, is part of a project that seeks to offer a metaphysical framework for practical markers of identity, and so represents an interesting attempt to provide a foundation for the ways we think about and identify with multiple selves. I argue for this conclusion by way of systematically canvassing the various considerations at play in the debate about the relation between selves in Kant. I begin with a bit about this debate.

There is widespread agreement that Kant must hold that there are two *kinds* of selves, one a so-called phenomenal, or empirical, or experiential self, and the other a so-called noumenal, or transcendental, or real self. The question, then, is whether these different kinds represent two or one distinct, countable entities. What is the relation between the empirical and the transcendental self? In broad strokes, three interpretive positions have emerged: (a) Kant is committed to there being one self, (b) Kant is *apparently* committed to there being two selves, but as interpreters we need to amalgamate selves, i.e. read some away or into others such that we end up with only one

self, (c) Kant is committed to there being two selves, but that is a strange or implausible view of personal identity.⁹⁰ Position (a) is motivated by Kant's practical philosophy and his theory of moral responsibility, (b), by his theory of self-consciousness, the ascription of conscious states, and the referent of the first-personal pronoun 'I', and (c) by the entailments of Kant's idealism and his theory of freedom and causality.

The picture that emerges from this is that these motivations are mutually inconsistent: an account of the relation between selves has to prioritize one motivation at the expense of the others. In taking this to be the case, the majority of readers have ended up proposing 'quick' solutions to the problem of the relation between selves in Kant: that is, adopting one of position (a)-(c) without explaining how that corresponds with the rest of the Kantian considerations, or arguing that achieving this correspondence means we must set aside (ignore or deflate) the real self in our accounts.

I argue, by contrast, that these three motivations not only *can* but *must* all hang together in our best account of the relation between selves in Kant: each is both textually and philosophically non-negotiable and so jointly circumscribe how we can model the relation between selves in Kant. Insofar as this is the case, there is no possible 'quick' account of this relation. We must consider

⁹⁰ For (a) see Adams 1997, 822, Marshall 2013, and Adickes 1924, 24, though the last is motivated by the considerations for (b), a position advanced by Kitcher 1984 and Van Cleve 1999, 182-6. Strawson 1966, too, could be considered a version of this view, insofar as he takes there to be two selves but no possible way to identify them. A commitment to there being two selves and an explanation of their relation (c) is suggested in Wood 1984 and Watkins 2005, 324-5, although both of their focus is on understanding the possibility of causality. It is explicitly held in Walker 2010, 836-7, and most clearly by Jauernig 2021, 285-294. I maintain that, despite this, these readers do take this to be an implausible account of identity because they move, when discussing either ethical or commonsense considerations, into the view that there is one self. E.g., Watkins' remarks on self-constitution (359) or Jauernig's commitment to a capital-S self (287).

specific models that can account for all three motivations and see how they fare in the context of Kant's system.⁹¹

This is what I do in this chapter. I begin (§1) with some preliminaries to specify what question we are asking when we ask about the number of selves, and why exactly this question gives rise to a puzzle. I then argue (§2) for three non-negotiable desiderata for understanding the relation between selves in Kant. In putting forward these desiderata, I deepen the original puzzle and thereby conclude that there is no quick solution available. In §3, I turn to consider two models for the relation that can be taken to meet these desiderata: what I call the Substance/Accident Model, and the Whole/Part Model. Both models are versions of accounts defended in the literature, and while there is value in both, I argue that both run into serious problems when we attempt to fully spell them out. Both, then, ultimately fail to properly solve our puzzle.

The main takeaway I draw from this section is that a model which attempts to preserve either strict or partial identity between the selves doesn't work. In §4, then, I propose a new model—what I call the Actor/Activity Model—that takes seriously that the real and empirical selves are distinct, and I consider the textual and philosophical merits of such a model. While I argue that this third model is the best candidate for explaining the relation, I do not intend this as a knock-down proof that this is the right way of characterizing the relation between selves for Kant. My

⁹¹ Marshall (2010) also goes through the different demands that weigh upon Kant's metaphysics of the self, and proposes a model (the "Kantian Effect-Relative Account") that he argues satisfies these demands. There, he does not consider the demands of morality or moral action, and from the paper where he does (Marshall 2013) it is not evident, at least to me, how the Effect-Relative Account could handle the considerations nicely articulated there. In any case, the strategy of the 2010 paper is similar to the one I undertake here and brings to light several important elements of Kant's metaphysics of the self. In particular, my work here is indebted to work done in that paper and in Jauernig (2021) arguing that Kant *does* have a metaphysics of the self, contrary to the long-standing anti-metaphysical and deflationary assumptions about the Kantian self.

goals are, first, to show that in jointly understanding the considerations that inform Kant's metaphysics of the self, there is only a small subset of explanations of the relation available to us and that these explanations rule out the straightforward assertion that there is one self. Second, I argue that understanding there to be genuinely two selves in Kant is not a problem nor nonsensical: this understanding can likewise satisfy all the desiderata for a plausible and productive account of the self.

§1. Clarifying the Main Question

This chapter investigates the question: how many selves does Kant think we have? More precisely, I am here asking what metaphysical relation holds between two kinds of entities that are describable as selves: is it numerical identity, or a relation that preserves distinctness? I begin with a brief clarification of each of the parts of this question. First, what I am picking out when I ask about the 'self'. Second, what it means to either identify or disambiguate these two kinds of selves. And third, why this question is not settled by appeal to a more general metaphysical relation. These three clarifications will give us the first-pass statement of Kant's puzzle of the self-relation.

§1.1. Kant's concepts of self, human being, and person

Here is a sample of the passages that prompt questions about the number of selves in Kant's account. I have set the most relevant parts in bold.

Space itself, however, together with time, and, with both, all appearances, are not things, but rather nothing but representations, and they cannot exist at all outside our mind; and **even the inner and sensible intuition of our mind (as an object of consciousness)**, the determinations of which through the succession of different states is represented in time, is not **the real self as it exists in itself**, or the transcendental subject, but only an appearance of this to us unknown being, which was given sensibly. (A492/B520)

Yet the human being, who is otherwise acquainted with the whole of nature solely through sense, knows himself also through pure apperception, and indeed in actions and inner determinations which cannot be accounted at all among impressions of sense; he is obviously is in **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." (A546/B574)

For, that a **thing in appearance** (belonging to the world of sense) is subject to certain laws from which **as a thing or a being in itself** it is independent contains not the least contradiction; that he must represent and think of himself in this twofold way, however, rests as regards the first on consciousness of **himself as an object affected through the senses** and as regards the second on consciousness **of himself as an intelligence**, that is, as independent of sensible impressions in the use of reason (hence as belonging to the world of understanding). (GM 4:457) (see also KpV 5:162, A554/B582, B429-31, B407, GM 4:452, Prol. 4:344, KpV 5:97-8, B155).

Kant sometimes refers to selves or the self (the substantive form *Selbst*), sometimes to objects (an affected object, a merely intelligible object, a thing in itself), sometimes to the mind, and sometimes to these things as features of the simple reflexive form (*selbst*, as in 'myself', 'himself'). For our purposes, all these concepts are co-extensive: they track the referent of the 'I', that thing that seems like the bearer of my traits, what I am able to ascribe or determine with unique or personal facts about me. For this reason, they are all properly called 'self', and I will refer to that that which satisfies this definition as 'self' throughout.

This concept of a self is distinguished from a broader concept which includes as an element the self or selves: the human being. Kant is clear that a human being, whether it is considered as an ontological or a physiological entity, is an *embodied* rational being: all of us only have one body, hence all of us are only one human being.⁹² The interesting question at hand, then, or the puzzle

⁹² See A546/B574, B409, B415, M L1 28:224, MM 29:862, 29:877, 29:878, A550/B578. In a sense, then, this brings Kant close to Locke, who also proposes distinct identity conditions for the 'human being' and the 'self' (see, e.g. *Essays* II, XXVII 6-9). Kant, though, is distinct from Locke's position insofar as he does not take these conditions to be entirely distinguishable: instead, the identity of the human being as a whole supervenes on the identity of a self as one of its

inherent in Kant's theory, is whether each *human being* has one or two *selves* that they identify with. My focus here is on the specific notion of self, rather than the broader conception of human being.

This target concept of the self has a lot in common with Kant's concept of person, but they are not equivalent. The self, for Kant, can be identified synchronically. Persons are selves that are identified diachronically, either with a specifically temporal criteria or by the attribution of certain actions to them. I will return to what makes a self a person on Kant's view in §2.3.

§1.2. Numerical identity and distinctness in the Kantian context

The above passages are all ambiguous as to whether Kant intends there to be one or two selves when that question is taken to be one about numerical identity (i.e., a question about tokens of selves). But clearly, the passages do appeal to two different *kinds*, or types, of things: (1) things-in-themselves, intelligible objects, or noumena, contrasted to (2) appearances, or empirical objects, or phenomena. For clarity, I will refer to the self that is of kind (1) as the 'real self', and the self that is of kind (2) as the 'empirical self'.

This distinction between kinds is the heart of Kant's philosophical system of transcendental idealism, which has two relevant core commitments. The first is to the ideality of space and time: what is empirical necessarily appears in space and time and so is characterized by spatio-temporal properties. What is real, on the other hand, is not in space and time (non-spatiotemporal). The second, related core commitment is to the *mind-dependency* of appearances.

components. As pointed out by Lu-Adler (2023, 12-3) and Allison (2011, 209-18), Kant does use the term 'humanity' to refer to the *rational nature* of human beings general, but not to the human being as such or any particular human.

No matter which of the various possible ways the relation between things in themselves and appearances and the specifics of mind-dependency are spelled out, this commitment appears in some form. Given this commitment, when we entertain the possibility that the real and the empirical selves are distinct, we cannot be seeking, in Cartesian terms, a “real distinction”: a distinction between two things that can subsist or exist independently of any other things.⁹³

These commitments give shape to the puzzle that is entrenched in understanding the relation between selves in Kant. Basically, they can be spelled out as compatible with both the distinctness of these two selves, and are compatible with the identity of these two selves. Here is what I mean. On the one hand, that Kantian selves cannot stand in a real distinction need not prevent us from thinking that they are distinct entities. Important to his idealist metaphysics is the claim that, despite the fact that appearances are not ontologically basic, they are nevertheless “empirically real”. All the details of empirical realism and its relation to transcendental idealism are far beyond the scope of this chapter, but its slogans satisfy here. Appearances qualify as the real objects of scientific study and can be the bearer of distinct properties and determinations. Insofar as the empirical self is an appearance, it too can be thought of empirical real in this way.⁹⁴

⁹³ Famously, Descartes contends that there is a real distinction between mind and body as distinct substances with irreconcilable primary properties (*Meditations*, 86). For Kant, appearances depend, at least partially, on human minds for their existence: so too does the appearing self. In other words, numerical identity cannot be determined by what can subsist without the other thing, because it will always be the case that an appearance is delimited at least partially by being mind-dependent. If we think that the real self has at least something to do with my mind, then we have an argument which concludes that the empirical self depends (again, at least partially) on the real self for its existence.

⁹⁴ This highlights one way to spell out a distinction between selves absent independent subsistence. If the objects of sciences (like psychology) are real objects, then Kant is surely at least entertaining the claim that the empirical self is (a) an object of scientific (psychological) study, and (b) that it is a distinct object of study from the real self (which cannot be an object of psychology, either rational or empirical). Though the self, and not the human being, is our target concept here, Kant’s clear and abiding interest in the psychological and especially anthropological study of the human being who lives in this, the actual spatiotemporal world, is one of his clearest examples of the scientific study of empirical entities. For further discussion of this, see Lu-Adler (2023, Part II). See also Kant’s claims in the *Metaphysical Foundations of Natural Science* (4:471) where he emphasizes that, although a *natural science* of inner sense is impossible

On the other hand, the mind-dependency of appearances would allow us to hold that two things which do not share all properties can nevertheless be thought to be identical (one entity) without violating the principle of the indiscernibility of identicals. The indiscernibility of identicals states that: if x and y are numerically identical, x instantiates a property if and only if y instantiates that property. If we take x to be the empirical self, then the empirical and real self clearly stand in violation of this principle: the empirical self instantiates the property of *being in time* while the real self does not. However, it is also traditionally thought that the indiscernibility of identicals does not scope over mind-dependent properties: if the discernable differences between two entities are *mind-dependent*, they can be taken to fall outside the scope of the principle.⁹⁵ Since the fact that the appearing self instantiates temporal properties depends on human minds conceiving of and intuiting it *as* an appearance, 'is temporal' could be taken as a mind-dependent property that falls outside the scope of the principle. In this way, the empirical and the real self could be taken to instantiate contradictory properties while still be taken to be numerically identical on the traditional understanding of numerical identity.

This gives us the first-pass statement of the puzzle that concerns us here.

(because inner objects do not remain fixed before observation and are not reducible to mathematics) nevertheless psychology, as "the empirical doctrine of the soul," is possible as "a historical doctrine of nature, and, as such, a natural doctrine of inner sense."

⁹⁵ E.g.: Haley believes that the morning star is blue. Haley does not believe that the evening star is blue. The morning star has the property *is believed by Haley to be blue*. The evening star does not have the property *is believed by Haley to be blue*. But nonetheless, the morning star and the evening star are numerically identical, just different names for the same entity. The reason that this numerical identity is consistent with my set of beliefs is because those contradictory properties are mind-dependent ones. Obviously, the mental relation is strengthened beyond 'belief' in the idealist case. This principle plays a large part in a central debate in Spinoza scholarship, viz. whether the mind and body are one and the same thing. See Della Rocca (2003) and discussion in Morrison (2022, 309-10).

The Self-Relation Puzzle, First-Pass: Kant appeals to two selves. Though these are clearly two different kinds of things, his metaphysics leaves room to conceive of these two kinds of things as identical, and it leaves room to conceive of them as distinct. How should we understand the relation between selves?

§1.3. Why an appeal to a general metaphysical relation is not the solution

This puzzle raises the question of why I am asking specifically about the relation between the empirical and transcendental self, and not the relation that obtains between appearances and things in themselves more generally. For some context, Kant scholarship has long been occupied with a debate over whether transcendental idealism is a *one-world*, or a *two-world* view. A one-world view claims that appearances and things in themselves are one thing, where we only interact with or perceive the very same thing in itself as it appears. A two-world view claims that appearances and things in themselves are distinct entities that either stand in some relation (usually a grounding relation) or float entirely free of each other. Call someone a committed x-worlder if they maintain that this principle holds consistently, i.e. that it is true for any appearance/thing in itself pairing without exception. It should follow that if you are a committed one-worlder, then there is only one self. Likewise, if you are a committed two-worlder, there are two selves.⁹⁶

The problem with running this as a ‘solution’ to the problem of the number of selves in Kant is that it doesn’t actually solve any of the parts of the outlined puzzle, especially—as we will see—

⁹⁶ There is a further problem that an account of the number of worlds is not well-defined. For discussion of why this question needs to be at most metaphorical, see Walker 2010. For a criticism about the precision of the terminology, see Jauernig (2021, 4-15). For examples of different positions, see Langton (1998), Allison (2004), Allais (2015), and Jauernig (2021). For the various different ways each of these positions can be described as a one- or two- world view, see Stang (2024). By contrast, a question about whether the empirical and the real self are distinct or one is well-defined.

as it becomes more complex in light of the considerations I canvass in §2. For now, I will just say that there is a real question of whether the self is an instance of the general relation between phenomena and noumena, or whether it is an exception to a general relation. If we say the first, there is a question about the special status that Kant affords rational beings. If we say the second, there is a question about why the self is an exceptional case.⁹⁷ I illustrate these complications to motivate my choice of focus: we should start and attempt to settle the identity relation between selves first, and then go back and integrate it into the general metaphysical picture and system. The question of the nature, relation, and quantities of the self presents an inextricable explanatory puzzle for which a solution does not simply fall out of a general relation.⁹⁸

§2. Desiderata for an Account of the Identity Relation Between Selves

To respond to this puzzle, we should look to what Kant has to say about the self: what considerations he is drawing upon, and what he is motivated by in talking about the self. I turn now to survey the demands that are made on a Kantian account of the self and identity. I formulate these demands as three distinct desiderata for an account of the relation that obtains between the real and the empirical self: what I refer to as the Self-Reference Desideratum (§2.1),

⁹⁷ This also speaks to why the question of how many selves should be the starting point for the question of the relation between appearances and things in themselves: while it is not evident to me (in that it could go either way) that the self plays a special *metaphysical* role in Kant's system (i.e. that the parts of the self stand in a unique relation, amongst all things-in-themselves/appearances), it is clear that it plays a special *epistemic* role: we have some insider information about the self and not about other objects. Whether or not the self should be taken as a general test case, or whether it is a special instance of a unique relation, when it comes to conceiving of the general metaphysical principle in which things in themselves and appearances relate, this special epistemic role does situate it as the way into conceiving of the relation between domains.

⁹⁸ Even further reason to focus in on the question of relation that obtains between selves in Kant is supplied by a strand of interpretation that holds that questions about the identity of appearances and things in themselves are contentless, *except for* in moral contexts, that is, except for in the case of the human agent. See, e.g., Allison 2004, n.19. While I do not agree with this claim about general contentless, it does provide even more reason to focus first on the self. I will come back to these considerations once I have further specified our puzzle, at the end of §2.

the Moral Action Desideratum (§2.2), and the Personality Desideratum (§2.3). Running through these desiderata will both clarify and deepen our original puzzle.

In presenting these desiderata, I argue that each of them is non-negotiable. I determine their status as non-negotiable insofar as they track the overlap of key positions in Kant and key elements of a philosophically functional account of the self. That is, they capture *both* bits of the Kantian theory that must be accommodated, *and* work that a theory of the self and identity should do in order to be properly explanatory. It is the force of this overlap that makes each of them fixed points on which to map an account of the relation between selves.⁹⁹

§2.1. The Self-Reference Desideratum

The first desideratum arises from Kant's account of the commitments of self-consciousness, the capacity for representation, and the nature of the 'I', and it comes from commonsense considerations of what we are looking for when seeking an account of the self.

This desideratum—what I call the Self-Reference Desideratum, or D1—circumscribes what kinds of things can be properly be called a 'self'. The desideratum is this: if something is a self, it can be the proper referent of the first personal pronoun expressed as 'I'. If some x cannot be the proper referent of the expression 'I', then it is not a self. The dictum of this desideratum is that there are no otiose selves. For something to count as a self properly, it must be such that you actually identify with it: that is, that it is included in the referent of 'I'. Canvassing this desideratum touches on several of the points I made in the first clarification of our question, namely, just what

⁹⁹ I do not mean to claim that these three desiderata are the *only* non-negotiable considerations. If further considerations that share this status—that is, that are both philosophically and textually key—are found, then they should be taken as desiderata and the models for the relation that obtains between selves will have to be evaluated in light of these additions.

Kant takes a 'self' to be. But in understanding his arguments for the possibility of self-reference, we get some telling information about how we should be conceiving of the self and possible relation between selves. This is what I'll look at here.

These arguments come out the commitments of his account of self-consciousness and the capacity for representation. Kant famously claims that "the I think must be able to accompany all my representations, for otherwise [...] representation would either be impossible or else at least would be nothing for me (B132)." He develops this claim—that a capacity to think precedes my capacity to have or make sense of representations—in terms of the *self*, and *self*-consciousness. A capacity to think is intrinsically a capacity to think of your *self* as a subject (intrinsically in that thinking presupposes a subject, not that a subject is occurrent in each thought). Accordingly, he writes:

The thought that these representations given in intuition all together belong **to me** means, accordingly, the same as that I unite them in a self-consciousness, or at least can unite them therein, and although it is itself not yet the consciousness of the synthesis of the representations, it still presupposed the possibility of the latter, i.e., only because I can comprehend their manifold in a consciousness do I call them all together **my** representations; for otherwise I would have as multicoloured, diverse a self [*Selbst*] as I have representations of which I am conscious. (B134)

Eliding some of the technical complexities of this account, the following thesis emerges: the self is the referent of 'I' (or 'my': any first-personal pronoun, actually expressed or not). This is because the ability to think each thought as *my* thought is what gives representations their unity, a unity that precedes their comprehensibility. And the unity of my thought *is* the unity of my self.

It is important to recognize immediately that a claim about the *unity* of the self should not be equated with a claim about either the *singularity* or the *existence* of the self. If I were to recognize a multitude of discrete, non-unified representations, I would have as many selves as the number

of those representations: the self's existence is not conditioned on the unity of thought, and neither need it be a singular thing. Instead, I take it that this is an argument for conceiving of the self as that thing that can function as the referent of 'I': the Self-Reference Desideratum.

That Kant's account of self-reference exhibits a demand for unity rather than singularity is what makes consistent his otherwise *prima facie* contradictory statements that there is an identical self in thought, and that the 'I' has multiple distinct referents. Compare these two passages:

I am therefore conscious of the identical self (selbst) in regard to the manifold of the representations that are given to me in an intuition because I call them all together my representations, which constitute one (B135, see also B138).

But how the I that I think is to differ from the I that intuits itself (for I can represent other kinds of intuition as at least possible) and yet be identical with the latter as the same subject, how therefore I can say that I as intelligence and thinking subject cognize myself as an object that is thought, insofar as I am also given to myself in intuition, only, like other phenomenon, not as I am for the understanding but rather as I appear to myself, this is no more and no less difficult than how I can be an object for myself in general and indeed one of intuition and inner perception. (B155-6, see also A117)

The first passage describes a consciousness of an *identical* self, while the second describes two different 'I's' along with *two different* selves. In understanding 'identical' to be invoking a certain kind of unity, we can understand them to be part of the same account. What Kant calls the 'identical self' is the unification necessary for self-reference. The thing that is actually the referent of 'I' can be more than one thing, as long as there is sufficient unity amongst the referents such that can be properly identified with.

Two important things, then, emerge from getting this Self-Reference Desideratum on the table. First, that a claim about the unity of the self or the referent of the 'I' need not be a claim about singularity, and, in fact, in some of Kant's arguments, we get a better result if it does not. And second, we can't have a 'self' that we don't identify with (that doesn't function as the referent of

'I'). If the 'real self' is such an entity, then properly speaking (by both commonsense and Kant's own lights) it does not count as a self and so the relation question is rendered obsolete. Our account of the relation, then, must be such that it does justice to this identification. Since Kant calls the real self a *self*, any account of the relation that we offer must be such that it can make sense of the real self as proper referent of the 'I'. Likewise with the empirical self.

As for non-negotiability, this desideratum arises out of key part of Kant's theory—his account of thought and the applicability of *a priori* truths to the sensible world—and it arises out of a key part of any investigation into questions about the self, viz. what are you referring to when you express 'I'? See, for instance, Parfit's identity puzzles about mind-fusion, or Locke's thought experiments about the continuity of self-consciousness.¹⁰⁰ From this, then, we get our first desiderata: the account of the relation between selves must account for how these selves are the proper referent of the 'I'. The *easiest* way to preserve the relevant identification is, of course, to say that the relation between selves is strict identity: there is exactly one self. But this is not the *only* way.

§2.2. The Moral Action Desideratum

The first desideratum came out of considerations about the nature and function of thinking in Kant's system. The second desideratum arises from considerations about the nature of action—in particular, from the nature and possibility of moral action. This desideratum is that an account

¹⁰⁰ Parfit (1971), Locke (II, XXVII-23). See also Lewis (1983)'s claim that the self *perdures*, not *endures*: his motivation there is that philosophical expansion on a commonsense conception of the self or identity demonstrates that intuitions about self-reference can be substantiated by a certain kind of *unity*, not just singularity. Lewis is claiming that there are only a multiplicity of selves diachronically—a four-dimensionalist view holds that, at any *one* time, we only have one self. I discuss this more in contrast to Kant in §4.5.

of the Kantian self must be able to account for the evaluability of moral actions. Call this the Moral Action Desideratum, or D2.

Throughout his mature philosophy, Kant is committed to three crucial claims about the nature of morality and the possibility of ascribing moral responsibility to actions. First, that morality presupposes the ability to act *freely*: the ability to initiate events or effects without antecedent or external cause.¹⁰¹ Second, that every event and effect that occurs in a natural, spatiotemporal order—like that of the world we experience—is causally determined: every cause is in turn the effect of an antecedent cause.¹⁰² Third, morality is *real*: we are justified in ascribing moral praise and moral blame to actions.

These three claims can be made consistent, Kant argues, by distinguishing between a realm of phenomena, organized in terms of a natural, causally determined order, and a realm of noumena that is ungoverned by such an order. Distinguishing between realms in this way safeguards the second claim while making space for the first: the *phenomenal* realm is a causally determined system, but freedom is possible in the *noumenal* realm. The reality of morality thus presupposes that certain actions in the phenomenal realm are effects of a noumenal cause.

Since moral actions are the actions of *people*, this argument for the possibility and nature of moral action is in effect (or contains) an argument for the nature of the human self. Kant makes this explicit in describing the account.

¹⁰¹ “By freedom in the cosmological sense, I understand the faculty of beginning a state from itself, the causality of which does not in turn stand under another cause determining it in time in accordance with the law of nature.” (A533/B561)

¹⁰² This includes human actions in the spatiotemporal world. Kant argues that if we had enough empirical information “there would be no human action that we could not predict with certainty.” (A550/B578)

In the antinomy of pure speculative reason there is a similar conflict between natural necessity and freedom in the causality of events in the world. It was resolved by showing that there is no true conflict if the events and even the world in which they occur are regarded (and they should also be so regarded) merely as appearances; for, one and the same acting being as *appearance* (even to his own inner sense) has a causality in the world of sense that always conforms to the mechanism of nature, but with respect to the same event, insofar as the acting person regards himself at the same time as *noumenon* (as pure intelligence, in his existence that cannot be temporally determined), he can contain a determining ground of that causality in accordance with laws of nature which is itself free from all laws of nature. (KpV 5:114)

Insofar as a 'being' both appears and is real, it can be predicated with both a mechanist causality and function as the ground of non-mechanist, free causality.¹⁰³ The possibility of freedom is located with the real self that grounds the morally evaluable actions of the empirical self.

A theory of the Kantian self must be able to explain how this is possible, and an explanation of the relation that obtains between the empirical and the real selves must be able to satisfactorily account for the demands on both sides of the equation. On the one hand, the account of the self needs to include reference or make recourse to the noumenal self as the determining ground of the causality necessary for morality. In order for moral evaluation to be apt or justified, those evaluated actions must originate in the supersensible realm of things-in-themselves. If the real self is not involved, then there is no basis for blaming the phenomenal self for their acts.

On the other hand, we don't want to prioritize the noumenal agent at the expense of a plausible account of evaluating empirical actions. Here is what I mean. We ascribe praise and blame to

¹⁰³ Several readers (Marshall 2013, 427, Adams 1997, 822) take passages like this as unequivocal evidence that Kant takes there to be one self, appealing to the language as "one and the same acting being". But as I argued in §2.1, Kant often uses language of 'oneness' for appeals to a unity, i.e. several things unified under a joint function, and so they do not necessarily indicate singularity. A further reason to think that there is no unequivocal reading of this passage is that it is almost identical to the account Kant offers in the Resolution to the Third Antinomy, except there he can be taken to clearly appeal to two *distinct* beings (see, e.g. A538-9/B566-7). Again, Kant's language leaves the question open: it is one that has to be settled by philosophical as opposed to purely textual considerations.

actions that occur in *this*, sensible world, and so to the actions of the phenomenal selves we encounter (including one's self). We—and Kant—don't want to lose our grip on this fact. But if the blame is placed entirely on the noumenal self *and* there is no discernable connection between the noumenal and phenomenal self, then there seems to be no basis for evaluating the empirical actions for which we *want* to hold people responsible. If I break my promise or kick you in the shins, we should want to condemn *that* act, of promise breaking or violence. If the noumenal self is entirely divorced from these actions and its connection to the empirical self a mystery, we get the very unintuitive result that my kick is not itself morally evaluable.¹⁰⁴

The upshot for my investigation here is that the relation between the selves needs to be characterized such that we get both the ground of free causality and the ability to hold agents properly responsible for empirical actions. The account we give of that relationship needs to explain how the actions of the real self have empirical consequences. The relation between the real and empirical self needs to be one that preserves transitivity. As with the first desideratum, the most *straightforward* way to meet this demand is to appeal to strict numerical identity, but it is not the only way. Again, this desideratum—the Moral Action Desideratum—is non-negotiable,

¹⁰⁴ Some might object that Kant does take the real morality of our actions to be inscrutable in this way, as suggested by A551/B579n. There are two rejoinders to that. First, that note is clearly an epistemic point that is distinct from the metaphysical claim: even if I cannot know with certainty or cognize the real motivation behind my actions, the metaphysics should still shake out that, as a matter of fact, actions in this world are praise or blameworthy. Second, Kant is not a fatalist about our moral motivations. In the *Religion* (6:46-8), he argues that a person can undergo genuine moral progress by changing their fundamental maxims, i.e., changing their intelligible character (the law that governs the causality of the self as a thing in itself). Since our intelligible character is not in time, this argument at least gestures at a reciprocal relation between the actions of the empirical self and the nature of the real self. Depending on how you read Kant's claim that "[human beings] who can assess themselves and the strength of their maxims only by the upper hand they gain over the senses in time", the process of moral progress is at minimum *reflected* in and perhaps more strongly *influenced* by empirical actions. This is all to say that the account we give of the self cannot render the ascription of moral responsibility in the empirical world inert, for reasons of philosophical plausibility and by Kant's own lights.

as it sits at the intersection of a key part of Kant (his theory of free, moral action), and a key part of a theory of the self (how we can be held morally accountable).

§2.3. *The Personality Desideratum*

While the previous two desiderata both seem to point us in the direction of thinking the relation between the real self and the empirical self should be strict identity, this last desideratum will show why that can't be the case, at least understood as such. I call this the Personality Desideratum, or D3. What is at the heart of this desideratum is a need to account for the differing identity and persistence conditions that Kant attributes to appearances and to things in themselves, and in particular to the differing ways that an appearing self can be identified and persist *as a person* and that a real self can be identified and persist *as a person*.

A helpful place to begin is with the discussion of Identity and Difference from a section of the *Critique of Pure Reason* known as the Amphiboly, wherein Kant diagnoses the errors of Leibniz's philosophy insofar as Leibniz took (as Kant thinks he did) objects in space and time for things in themselves. The first error he targets is Leibniz's principle of the identity of indiscernibles. Kant writes:

If an object is presented to us several times, but always with the same inner determinations, then it is always exactly the same if it counts as an object of pure understanding, not many but only one thing, but if it is an 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 places of these appearances at the same time is still an adequate ground for the numerical difference of the object (of the senses) itself. Thus in the case of two drops of water one can completely abstract from all inner difference (of quality and quantity), and it is enough that they be intuited in different places at the same time in order for them to be held to be numerically different. [...] multiplicity and numerical difference are already given by space itself as the condition of outer appearances. (A263-4/B319-20)

While I began by arguing that there is no general reason to deny the identity of things in themselves and appearances, we now get a particular account of why they should be taken to come apart. This passage offers two distinct identity criteria: the identity of indiscernibles (if two things have all the same inner determinations, they are identical), and a principle of spatiotemporal individuation (if two things have all the same inner determinations but occupy different spatial locations, they are distinct—or, more simply, spatial location can differentiate).

It is clear from this passage that neither principle applies *across* categories: that is to say, each is restricted within a domain, either the world of appearances or the world of things in themselves. The passage also strongly suggests that each principle *does apply* within their domains: that is, we *can* individuate appearances based on spatial properties and determinations, and we *can* (or, properly speaking, we *could*, with access to more information) individuate things in themselves based on their intrinsic properties. From the Amphiboly alone, there is some reason to be hesitant about subscribing to such a claim—after all, Kant is there commenting on Leibniz, and therefore not necessarily committing himself to the truth of those claims.

I argue, though, that we can find these general considerations reiterated in the specific case of the self. In particular, Kant brings these different identity criteria into play when considering how we can identify a *self* as a *person*. In these cases, he is evidently making a positive claim, substantiating the assertion in the Amphiboly that the identity conditions of an appearance are essentially spatiotemporal, and since things in themselves are essentially *nonspatiotemporal*, the identity conditions for a thing in itself must come apart from those of an appearance.

Consider what Kant offers as a definition of “personality” in the Paralogisms. The Paralogisms is Kant’s rejection of the *inferences*, but not the definitions, that lead the “rational psychologists” to conclude that they have a priori access to different determinations of the self or the soul. This is the definition given there of personal identity: “What is conscious of the numerical identity of its Self in different times, is to that extent a person.” (A361).

Such a definition or account of how we identify a self as a person is clearly a temporal one. You are a person if and so long as you are conscious of the sameness of your self in *different times*. Contrast this account of what it is to be a person with another definition Kant offers in his *Metaphysics of Morals*.

A person is a subject whose actions can be imputed to them. Moral personality is therefore nothing other than the freedom of a rational being under moral laws (whereas psychological personality is merely the ability to be conscious of one’s identity in different conditions of one’s existence). (MM 6:223)

The definition offered here (and explicitly contrasted with another definition, i.e. the temporal criterion found in the Paralogisms) is that you are a person if your actions can be imputed to your self. As we saw in the previous section, a self or subject whose actions can be properly imputed to them—a self that is free—can only be one that is not in space and time (and so one that exists outside of the deterministic causal order of nature).

Just like in the Amphiboly, then, we got two distinct identity criteria. What it is to be a person is have a capacity to produce actions versus what it is to be a person is to be reidentified across time.

Taking the Amphiboly distinction together with Kant’s remarks about a personality criterion that is indexed to time vs. a personality criterion that can only be applied to things in themselves give us good grounds, I argue, to conclude that Kant takes the empirical self and the real self to have

differing identity and persistence conditions. This gives us our final desideratum: an account of the relation between selves must account for their different identity or persistence conditions. As with the other two, this is both textually and philosophically non-negotiable. Textually, because it arises out of the way Kant argues space and time feature in differentiating objects or subjects. Philosophically, because it concerns a key question for an account of the self: how that account relates to what it means to be a person, and why a self is not just the same as the person.

I have now argued that there are three desiderata for an account of the relation between selves in Kant. The Self-Reference Desideratum (D1): an account of the relation that obtains between selves must accommodate how each self can be the proper referent of the 'I'. The Moral Action Desideratum (D2): an account of the relation must explain how moral action is possible. The Personality Desideratum (D3): an account of the relation must allow for the different identity/persistence conditions for each self. I have argued, too, that each of these desiderata are *non-negotiable*: they are born out of key parts of Kant's texts and out of key elements of a theory of the self (how we think, how we act, and who we are as people). If we wish to give an account of the self in Kant that does justice both to the *self* part and to the *Kant* part, that account should be indispensably circumscribed by these three desiderata. Going forward, I will refer to these as D1, D2, and D3.

With this commitment in hand, we can return to and restate our original puzzle. On a first-pass presentation, the puzzle was simple: does Kant's philosophy countenance one or two selves? Now we have the material in hand to redescribe this puzzle as follows.

The Self-Relation Puzzle: How can we understand the relation that obtains between the real and the empirical self such that this relation preserves the unity required to satisfy (D1), the transitivity required for (D2), and the distinctness required to satisfy (D3)?

It is possible solutions to this puzzle that I turn to now.

§3. Evaluating Two Proposed Models for the Understanding the Relation

Interpreters frequently elide the complexities of the relationship between the real and empirical self, asserting that in the case of the self—and often only in the case of the self—it is clear that there is one object. See, for instance:

When practical reason is allowed into the discussion, it is clear that Kant postulates both noumena that definitely are identical with some phenomena and noumena that definitely are not. We ourselves are the prime example of the former. (Adams 1997, 822)

A number of passages concerning these phenomena do seem to require the concept of a *single entity* which may be regarded both phenomenally and as a thing in itself. Furthermore, it might be argued, if such were not Kant's view, then Kant's attempt to employ the phenomenal/noumenal distinction in order to account for certain facts of moral experience would fail. (Aquila 1979, 305)

As Kant describes it our distinctive status hinges on our being both phenomenal and noumenal. So, if we take him at his word, it is crucial to Kant's philosophical project that the self is a single thing with both noumenal and phenomenal sides. (Marshall 2013, 427)

In so stating, scholars limit their focus to what I have called D2 (the moral action desideratum).

And while, as we have seen, they are all clearly tracking the most straightforward way to ascribe moral responsibility to the actions of an embodied, empirical agent, a thorough examination of both this consideration in itself and in the context of the various points Kant marshals to describe and discuss the self make evident that it is far from clear that Kant wants to assert that the real and the empirical self are identical. First, what appear as *prima facie* appeals to singularity in Kant should instead be read as appeals to *unity*, where that unity need not be understood as

identity, including in the moral context. And second, this is a good thing, because strict identity is inconsistent with the distinct persistence conditions for the real and empirical selves.

My survey of this desiderata, too, makes evident that (pace Kitcher 1984, 138) the “focus of philosophical interest” is on *both* the empirical and the real selves, not solely the former: we cannot ‘solve’ our puzzle by cordoning off one side of the relata from our investigation. And it showcases why we should be suspicious of readings which deflate the real self into a thought entity of some kind: a mere fiction, a heuristic of reason, or a normative prescription for the development of the experienced self.¹⁰⁵ For if the real self was a thought entity, it is hard to see how it can ground free action in the way Kant’s moral theory requires it to.

The puzzle of the self-relation in Kant, then, eludes these straightforward answers: the content of each desideratum and the way all three interact narrowly circumscribe the possible explanations we can offer for this relation.¹⁰⁶ We cannot take the route of the majority of interpreters, and just *assert* that there must be one self with no real explanation of how that assertion fits with all of Kant’s motivations. In this section, then, I will propose and evaluate two models for this relation that *can* be made to satisfy all three desiderata, at least minimally. In considering these two models I will be drawing on some views from recent literature. Several of the authors who put forward versions of these models have not explicitly considered, as I do here, how these models

¹⁰⁵ See Kraus (2020), Dyck (2006), Allison 2004.

¹⁰⁶ It remains an open possibility that there are either (a) further non-negotiable desiderata which would further circumscribe the kinds of models we can entertain for the account, or (b) that there are other models that do satisfy all three desiderata as I have specified them. My focus here is on bringing our attention to the considerations and complexities that instruct answering this question in Kant. If (a), then the three models I propose will have to be reconsidered in light of those further desiderata, and if (b) then those further models will have to be evaluated and considered alongside the ones I go over in the remainder of the chapter.

can satisfy all the different motivations Kant invokes for the relation between selves. Nevertheless, as I will show, versions of these models can be made to fit our desiderata. But, in so doing—in understanding them to clear this bar—each model, I argue, runs into serious problems when we attempt to spell it out and integrate it into Kant’s system. Neither, then, can be the way we make sense of the relation between selves in Kant.

One last thing to say before I turn to these models. We have now deepened our puzzle, but our original question still stands: does Kant think we have one, or two selves? My claim is that in finding a model that is the best fit with this set of desiderata, we arrive at an answer to this original question. In rejecting these first two models, I will also be focusing in on an answer.

§3.1. Model 1: Substance/Accident (Unity of Inherence)

As a reminder, the puzzle that needs to be resolved is that the first two desiderata demand that that relation be an extremely tight one: one that preserves the unity of self-reference and the transitivity of moral action. The third desideratum aspires for distinctness: two selves that can come apart in their individuation and their persistence.

The first model I will consider is the model of substance to accident. Alternatively, we can cache this out as a model wherein the class of unity is the unity of *inherence*. The crucial feature of this model as applied to our case here is that the empirical self *inheres* in the real self, as a property (or set of properties), an accident, or a mode. Notably, this is model is one on which there is *one* self,

viz. there is identity of the real and the empirical self.¹⁰⁷ I call this understanding of the relation Model 1.

As a general account of things in themselves, a similar model is defended by Rae Langton (1998) and by Lucy Allais (2015), though neither indicate how this model could accommodate the case of the self. As an account of the self, a similar model is suggested by James Van Cleve (1999): see his claim that, “in being aware of the states of one’s noumenal self, one is not being aware of a second self any more than in being aware of the qualities of an apple one is being aware of a second piece of fruit”.¹⁰⁸

Here is how Model 1 can satisfy our desiderata. As the Van Cleve quote indicates, this model posits only one self and so that one self functions, in accordance with the demands of D1, as the referent of the ‘I’. When I speak of ‘I’, I can be referring solely to the properties of this self, but I am still referring to this one self.¹⁰⁹ On the substance/accident model, the real self (the substance) and the empirical self (the mode) can both be taken to be a subject, the former in the proper sense and the latter in the sense of the subject of thoughts and judgements. Nevertheless, they are numerically identical in the sense that referring to the properties of an x is also referring to the x (where x, in this case, is the self).

¹⁰⁷ Or, more carefully, the real and empirical self are actually only one entity: we would not want to say that I am numerically identical to all of my properties or state. By numerical identity here I mean to pick out this claim about oneness, referring back to question about whether the selves are distinct tokens as well as distinct types.

¹⁰⁸ Van Cleve (1999), 186 n.37.

¹⁰⁹ This is compatible with an understanding of the empirical (but not the real) self as the *subject* of our particular judgements (of the form x is f). The empirical self’s existence consists in the real self having certain predicates, but these predicates need not include ‘being the thing that is the bearer of empirical properties. In that way, it is consistent to say that the empirical self is not a real *substance*, while still allowing that it is a proper *subject*.

D2 is likewise satisfied: echoing the vast majority of commentators, it is one and the same self that is responsible for actions and that experiences the consequences of the actions.

And this model can accommodate D3. If the empirical self is just an accident or property of the real self, then they have can different persistence conditions while still constituting one thing. The red of the apple can fade into a green without the apple going out of existence: the apple persists while one of its properties does not.

As an added benefit, the substance/accident model can straightforwardly accommodate the language of 'one and the same' that we see Kant occasionally employ. While I have argued that this language does not in and of itself commit Kant to singularity or numerical identity, it is still the most nature reading of the expression. On this model, we can lean into this natural reading: the real and empirical selves are one singular entity.

The substance/accident model can satisfy our three desiderata, and this satisfaction brings along with it some nice textual clarity. So what are the problems?

There is one preliminary issue worth mentioning, although it is independently surmountable. In the essay 'What does it mean to orient oneself in thinking?', Kant claims there is something incomprehensible about the Spinozistic view of the psychological subject's being an accident (8:143n). For all I have been trying to make it comprehensible to us, then, Kant himself seems to think that a self can't be an accident, and so we should be suspicious about attributing this to him.

This suspicion finds purchase in what I take to be two main problems with this model. The first is that it significantly minimizes the role of the empirical self. The second is that it requires a

commitment both to the existence of real substances and to the identity of real substances and their properties: these commitments problematize several other of Kant's claims.

Problem 1: Minimizing the Empirical Self

For this model to work, the empirical self cannot be an *essential* property of the real self. If it were essential, then as it changed (and especially as it went out of existence) so too would be the real self. But as we have already seen, this is not the case: the persistence and identity conditions for the real self and the empirical self come apart. We are left, on this model, to hold that the empirical self is just *a* property of the real self.

This greatly reduces the role and importance of the empirical self, and so is hard to square with many of Kant's claims about, for instance, our self-cognition and our empirical desires. It is not *impossible* to square: there is no contradiction. This is a picture that minimizes both the significance and the *reality* of the empirical self: the empirical self is just the qualities of the real self, like the redness of an apple. But Kant seems to want to conceive of the empirical self as more than a mere state. While this may be a fine metaphysical model, there is reason to be wary about its appeal as a psychological model.

We can solve this problem by making the empirical self the (or a) *essential* property of the real self, but then we have lost our grip on how the model satisfies D3: the two selves cannot have distinct persistence conditions if one is essential to the other. If we commit to Model 1, either we minimize the role of the appearing self, or this model cannot satisfy all the desiderata.

Problem 2: Reconciling the Required Metaphysical Commitments with Kantian Humility

Even if one is comfortable with accepting such a minimal understanding of the empirical self, we encounter problems with this model insofar as we consider the consequences of the metaphysical commitments required to make this picture work, especially in light of Kant's epistemic strictures. In his mature writings, Kant is famously committed to a thesis about epistemic humility: there are limits on what human minds can cognize, and this cognition cannot extend to things as they are in themselves. There is a great deal of controversy over the consequences of epistemic humility and the non-cognizability of things in themselves.¹¹⁰ But it is evident at minimum that certain of his arguments either depend on or conclude that we lack any direct awareness of the determinations of things as they really are. I contend that the two crucial elements of the substance/accident model—that the real self is a substance, and that substances are identical with their properties—problematize Kant's claims about our epistemic strictures.

In light of Kant's epistemic restrictions, a dilemma seems to arise for Model 1. Either we say that the real self is an ontologically basic substance, or we say that we are unsure that the real self is an ontologically basic substance. The agnosticism of the latter horn is based both on Kant's general epistemic constraints, and on suggestions (e.g. at A803/B831) that for all we know, we could turn out to be modes at the noumenal level. But if we take this horn we are in a precarious scenario wherein the real self is *real* but not basic, making the empirical self the representation of an accident that belongs to something of which we know nothing (let's call this the mode of a

¹¹⁰ I am wading into this controversy here in even suggesting models for the relation between the real and empirical self, and especially in asserting the existence of the real self. I hope to have offered a great deal of support for the claim that Kant *does* appeal to the real self, and so there *is* a need to understand this relation. And as to the first, my objection to Model 1 is not *that* the model purports to explain the relation, but that the main tenets of this particular explanation problematize Kant's claims about our epistemic strictures.

mode problem). To avoid the precarity and potential infinite chain of this kind of scenario, we should take the first horn and say that the real self is a substance. But if we arrive at the fact that the real self *must* be a substance via the fact that the appearing self is an accident, and accidents require substances in which they inhere, then it seems like we can know that the self is a substance through introspection on the nature of our self as the subject of judgment, which is the inference structure Kant denies and dismantles in the Paralogisms (B410).

That is one worry about reconciling the metaphysics with the epistemology on this model. The other is this. If we are to preserve the identity reading of the substance/accident model (which is its main benefit), then we have to say that the empirical self is the extrinsic predicate or property *of the* real self. But that means that insofar as I am making judgements about my empirical self, I am making a judgement about the real self (that substance with the extrinsic property of being that empirical self). This brings at least one real substance—the self—into the world of space and time, which is at least potentially a problem for Kant’s epistemic strictures.¹¹¹

For these reasons, I argue that this model, despite meeting the minimum explanatory barrier, is not the right way to understand the relation. For our purposes here, then, this also rules out attributing a one-self picture to Kant: the relation between the real and empirical selves is not *strict* numerical identity. On to the next model.

¹¹¹ Although see Chignell (2022) for a view that argues that this is not a problem for Kant in at least certain cases, including the case of the self.

§3.2. Model 2: Whole/Parts (Unity of Composition)

The second model I turn to is a whole to parts model, or a relation that is characterized by the unity of *composition*. On such a way of understanding the relation, the real and empirical selves are distinct parts that together form a composite whole. Qua parts, they are distinct: this, unlike the first model we canvassed, is a two-entity picture. But qua elements of a composite, they stand in a very tight relation. I will call such an account of the relation Model 2.

Variations of this proposal dominate considered diagnoses of the relation between the empirical and real self in Kant. Ralph Walker (2010) and Anja Jauernig (2021) both propose that the real and empirical self are independent entities that jointly compose *some* thing: the self properly understood. The biggest upshot of Model 2, if it can be proposed successfully, is that we can have both identity and distinctness: we can be committed to two distinct entities that are called ‘selves’, while also granting that properly speaking there is only one ‘self’, thus reconciling all the full-bore Kantian metaphysics with the intuitive claim I began with.

One of the problems that I am going to raise for this model is that it is not clear whether we should take both the individual parts as well as the whole they compose to be selves¹¹²: that lack of clarity is reflected in thinking through how the model speaks to the identified desiderata. Nevertheless, I think there is a way we can understand Model 2 to at least *prima facie* satisfy our desiderata.

¹¹² Both Jauernig and Marshall (2010) refer to both respective parts and the composed whole as the ‘self’ at various points. Jauernig does this intentionally, reserving ‘Self’ for the whole. I take it that Marshall is arguing (e.g. 14-5) that it is only the whole that is properly speaking the self (it is important to note, too, that he is not arguing for a whole/part picture, but what he refers to as an “effect-relative” composition relation); he does not speak to the fact that the entities he takes to compose this proper self are themselves described as selves.

For D1, we have two options. We either say that (1) the referent of 'I' is the composite, of which the two selves make up distinct parts, or (2) that there are two distinct referents of two distinct instances of 'I', but these come together to compose the proper/whole self. Both readings, but especially the second, can speak to this crucial part of Kant wherein he claims that the 'I' is doubled and yet identifiable as the same subject.

For D2, we should want to say that the whole (the 'proper', composed self) is what is held morally responsible, with the possibility of that responsibility being grounded in the joint action or properties, perhaps, of both its parts—that is, that the real self provides the ground for action and the empirical self the ground for the effects or results of that action.

Finally, D3. As distinct parts, the real and empirical self can have distinct persistence and individuations conditions. The persistence of the whole is derivative on the joint persistence of both its parts: once one part is no longer in existence, the whole too no longer exists.

As a bonus, although I grant that an identity reading of claims like 'x and y are one and the same' is a more natural one, I have argued that Kant himself seems to be committed that 'sameness' or 'oneness' comes from unity, not from singularity. The whole/parts reading of the relation between selves nicely captures this interesting (and, I think, underappreciated) strand of Kant.

What, then, are the problems with this model? I will again focus on two main problems. The first is a worry about whether we can actually specify the nature of the composition relation that obtains between selves. The second is a worry about this picture yielding yet another self that we need to contend with.

Problem 1: The Nature of the Composition Relation

A first problem for Model 2 arises from thinking about just how exactly we should spell out the nature of the composition relation that obtains between the real and the empirical self: that is, in virtue of what do these parts form a whole? Both Jauernig and Walker gesture at an analogy with Descartes' 'intimate union' between the body and the mind, wherein two distinct substances function together as a composite human being. But not only are they both quick to stress that the analogy is limited—that is, that the relation that obtains between the Cartesian mind and body cannot be extended to the Kant case—it is also the case that the union between mind and body is one of the more mysterious parts of Descartes's theory. An appeal to Cartesian mind/body unity here seems to be clarifying by way of obfuscating. But if the composition relation between the real and the empirical self is *not* the same as the relation between mind and body, then how should it be understood?

This is a problem, and not just an open question, because when push comes to shove it is very difficult to see just how we *could* specify such a composition relation. There are several ways to make the point that an appeal to a composition relation between the real and the empirical self can only ever be vague or metaphorical. Two worries have already been raised by commentators. One, from Marshall (2013, 435) is how composition can obtain between two such metaphysically dissimilar things, especially one thing that has spatial and temporal properties (and so is quantifiable) and another that does not (and so is not). Another, from Walker (2010, 825), is about appropriate sortals: what does it mean to say *x* is the same *A* as *y* in this context? It cannot be to say that *x* is part of the same human being as *y*, because that is circular (that is, they both belong to the sortal concept 'human being'). But given the distinctness in properties and identity

criterion, it becomes difficult to sort *x* and *y* in the way required for partial or relative identity, and to quantify over these parts such that we can make sense of them being of the same whole.

I think the clearest way to push this as a distinctly Kantian worry is in trying to make sense of the mereological principles involved in the whole-part relation of the self. Kant has a lot to say about mereological structure. Relevant here is that he distinguishes between real and ideal wholes, arguing that a real whole is dependent on its parts, whereas for ideal wholes the parts depend on the whole.¹¹³ It is unclear how these varying structures can be accommodated in composition *across* domains. Is the Composed Self a real, or an ideal whole? If ideal, then the whole precedes its parts, but since one of the parts is *real* than this makes a real part dependent on an ideal whole, which gets the order backwards. If the whole is real, then we get an equally strange chain whereby a real whole depends on an ideal part, which is itself dependent on something real.

Along any of these lines, it remains unclear whether a defender of the whole/part account model of the self-relation could actually explain or specify how we get composition.

Problem 2: The Identity of the Whole and its Parts

The second problem is distinct but related. In trying to articulate ways to characterize the composition relation between selves, I have been speaking of the ‘composed self’ as the whole. But this may not be the best way to understand the model. An open question remains: is the unity or composition of the two selves supposed to generate a *third*, distinct self, or is the composed

¹¹³ See Mrongovius 29:851 for the distinction between real and ideal wholes. See Hogan (2021) and Heide (2020) for arguments for the significance of these mereological claims for Kant’s idealism and his views on space.

whole in fact the self (so that we only have one self, properly speaking)? And again, attempts to settle this question quickly result in problems for the model.

One reason to say that the whole (the composition of the real and the empirical self) is something above and beyond the sum of its parts is found in attempting to spell out the properties of the whole. The whole can't have properties that are identical with only one of its parts (that violates symmetry without reason). And it can't have properties that are identical with *both* its parts because the parts, in this case, have contradictory properties (temporal/non-temporal). If we are not to run into the contradictory properties problem, then we must say that the whole is characterized entirely by emergent properties. But in that case, it seems that we have just identified a *third thing* that we wish to call a self.

If we say that the whole is just identical with the parts that compose it (like how a 1lb weight is identical to the two .5lb weights that make it up), then the model can no longer satisfy D3. In order to preserve D3, the parts have to be able to change independently while preserving the whole, which is only possible if they are distinct from the whole. We should instead be committed to the distinctness of each part and the whole they compose.

I began with a puzzle that is generated because Kant refers to *two* distinct kinds of things as 'selves' or 'the self'. Model 2 takes each of these selves to be parts that compose (as per the previous arguments) a distinct whole. Taking seriously that Kant is not speaking loosely or facetiously when he refers to these selves, Model 2 has the strange—and worrying—consequence that there are actually *three* selves. It is worrying for two reasons. First, because it complicates our original question and the original worry about the overdetermination or overcounting of selves.

Second, because it seems like this third self (the composite) is the one that we should really be interested in, and that interest is hard to square with the way Kant writes in the myriad of passages that motivate the self-relation puzzle in the first place.

The virtue of Model 2 was supposed to be that it accommodated both identity and distinctness. But it seems in trying to do so we have actually just complicated our original question. Now we are asking, does Kant think we have one self, two selves, or three selves? Given this over-iterating of selves, and given that it is difficult to see how we could give more than a metaphorical gloss on how the real and empirical self actually do compose a whole, I submit that Model 2 likewise faces too many problems to be the best way of characterizing the relation.

§4. A Positive Proposal: The Actor/Activity Model

Let's take stock of what we have seen so far. I collected three desiderata for an account of the relation between selves in Kant: what I called the Self-Reference Desideratum, the Moral Action Desideratum, and the Personality Desideratum. In light of these three desiderata, we were able to rule out a whole swath of accounts of the self-relation: accounts that hold that the real and empirical self must be identical without explanation, accounts that hold that the real self is so removed from the picture that it need not figure in our best account, or accounts that hold that the real self is merely a thought entity.

I've canvassed two possible models for the relation that can be made to satisfy these desiderata, at least on the face of it. Model 1 explains the relation as that of substance to accident or property. Model 2 explains the relation as that of whole to part. While both models have their defenders in the literature, they both run into serious problems. A common feature of all the problems faced

by both of these models is that they arise because the model purports to preserve some sort of numerical identity of the selves, either strict or partial identity.

In this section, then, I will propose and evaluate a model that takes seriously that these two selves—the real and the empirical self—are distinct. They stand in an intimate relation to one another but do not ultimately compose, at least *qua* self, *one thing*. This model envisions the selves as standing in a relation of *Actor* to *Activity*, where the real self is the actor and the empirical self is its produced, characteristic activity. I'll call this Model 3. I argue that Model 3 is the best fit for all the Kantian considerations and has a great deal of explanatory power when applied to his writings. It does, though, involve denying the commonplace with which I began the chapter: denying, that is, that we only have one self at any given time. Here, then, I'll argue for the value of the model in Kant, and conclude with some considerations about why we might actually want to deny the synchronic singularity of self both for Kant and for a wholesale theory of the self.

§4.1. The Actor/Activity Model (Unity of Order)

Here is how we should understand the relation between the real and the empirical self on Model 3. The real self to be the actor or agent, that is undertaking and then sustaining a specific and characteristic activity. That activity itself *is* the empirical self: the events and states that, unified as process, deserve the name empirical self. So understood, the two selves are distinct, in the same way that I am distinct from my act of walking. This model does not advocate for any kind of composition in explaining the self: there is a self that is the actor, and there is a self that is the activity. And yet these two things are tightly unified: the existence of the activity, while distinguishable from, is dependent on being sustained by the actor. And likewise, the actor is informed by the activity they are producing: in order to *be* an actor, they must be producing an

activity. Insofar as there is unity between actor and activity, it is a looser unity than the partial or strict identity that comes from the unity of inherence or composition. But there is unity nonetheless, like the unity that adheres between a direct cause and its produced effect.¹¹⁴

§4.2. Textual Evidence for Model 3

I'll go over how this model satisfies the desiderata below. First, here is some textual evidence that Kant just does think of the selves in this way. There are several places in Kant where he seems to assert *both* that the self is *producing* activity, and that the self (or the 'I') is an act or activity. Here are two such passages.

An indeterminate perception here signifies **only something real**, which was given, and indeed only to thinking in general, thus not as appearance, and also not as a thing in itself (a noumenon), but **rather as something that in fact exists and is indicated as an existing thing in the proposition "I think."** For it is to be noted that if I have called the proposition "I think" an empirical proposition, I would not say by this that the I in this proposition is an empirical representation; for it is rather purely intellectual, because it belongs to nature in general. Only without any empirical representation, which provides the material for thinking, the **act I think** would not take place, and the empirical is only the condition of the application, or use, of the pure intellectual faculty. (B423n)

The **I think expresses the act of determining my existence**. [...] Now I do not have yet another self-intuition, which would give the determining in me, of the spontaneity of which I am only conscious, even before the act of determination, in the same way as time gives that which is to be determined, thus **I cannot determine my existence as that of a self-active being**, rather I merely represent the spontaneity of my thought, i.e., of the determining, and my existence always remains only sensibly determinable, i.e., determinable as the existence of an appearance. **Yet this spontaneity is the reason I call myself an intelligence**. (B157-8n)¹¹⁵

In both passages, Kant identifies the self, or my existence, or the 'I' alternatively with an *act* (as in 'the act I think', or 'the I think expresses the act of determining my existence'), while proposing that we are simultaneously aware of something *real* that is capable of *producing* acting (my

¹¹⁴ Another way to describe the unity here would be as the unity of *order*, like the unity that obtains between a general and their commanding army. For more on this and further kinds of unity, see Pasnau (2011, 556).

¹¹⁵ See also some descriptions of the mind and its functions in the *Critique of Judgement*, e.g. 20:220 and 5:254-255.

‘existence as a self-active being’). Model 3 offers a neat way to make sense of these different descriptions, and is borne out in Kant’s own language.

This model can also capture the causal or grounding language that Kant uses when talking about the relation between selves in the Resolution to the Third Antinomy.¹¹⁶ For example:

Thus every human being has an empirical character for his power of choice, which is nothing other than a certain causality of his reason, insofar as its effects in appearance this reason exhibits a rule [...]. Thus in regard to this empirical character there is no freedom, and according to this character we can consider the human being solely by observing, and, as happens in anthropology, by trying to investigate the moving causes of his actions physiologically. But if we consider the very same actions in relation to reason, not, to be sure, in relation to speculative reason, in order to explain them as regards their origin, but insofar as reason is the producing cause of them by themselves—then we find a rule and order that is entirely other than the natural order. [...] The empirical character is once again determined in the intelligible character (in the mode of thought). We are not acquainted with the latter, but it is indicated through appearances, which really give only the mode of sense (the empirical character) through empirical cognition. (A550-1/B578-9)

Here, Kant explicitly says that the empirical character of a human being is caused by reason, or the intelligible character. Since an action is traditionally construed as an event that is caused by a specific agent, understanding the real self as an actor fits with Kant’s references to causation.¹¹⁷

¹¹⁶ This language is what informs both Wood (1984) and Watkins (2005, chap. 5) in their respective accounts of how Kant takes human freedom to be possible. The focus of both authors is the free causality at issue—that is, they are primarily concerned with articulating how it can be the case that actions in the empirical world are described as free because they have a non-sensible origin, as opposed to spelling out the relation between the real and empirical selves. Nevertheless, both explain this via a feature of the relationship between selves, and in so doing both articulate accounts that I take to be close to my Actor/Activity model. One key difference may be that they would not want to describe the empirical self *as* an activity (rather, a substance or actor that itself acts in accordance with a noumenal choice). Another key difference is that both, especially Wood, focus on *choice* as the crucial explanatory element in both Kantian freedom and the relationship between selves. On my view, although I do not defend this here, the concept of absolute spontaneity at issue is the concept of an uncaused action, *not* a choice between options. This is how Kant can take it that both God and human beings instantiate this property, even though God does not choose when exercising spontaneity (see Rel 5:50n.). Accounting for these important differences, though, I am broadly sympathetic to the conclusions drawn by both Wood and Watkins and my account here can be read as at least in part as an attempt to specify and clarify the metaphysics of the self that underwrite these conclusions.

¹¹⁷ See Anscombe (1957), Ruben (2018). Despite this causal language, I do think it is important to stick with the language of actor and activity for this model as opposed to cause and event. The language of ‘activity’ is useful in capturing that the experience of this self is an on-going *process*, as opposed to a fixed event with a determinate individuation principle and spatiotemporal location. See Charles (2018) for an argument that the ontology of actions should be understood as

The ‘actor’ part of Model 3 accords well with Kant’s causal language. Likewise, the ‘activity’ part fits nicely with Kant’s varied claims about whether or not the *empirical* self—that is, the self in inner sense—can be cognized. I appealed in chapter 2 to a passage (B156) where Kant seems to claim that we can cognize the self as it appears, at least in principle, but he is not always consistent about this claim.¹¹⁸ How exactly the empirical self can constitute a fixed object such that it can qualify as an object of cognition is a disputed and unsettled question.¹¹⁹ Applying the Actor/Activity model offers a neat explanation of why Kant seems to vacillate on this point. The empirical self is an activity understood as an ongoing process. Qua process, it does have a principle of spatiotemporal individuation, but this is not a fixed principle. It *could* be the object of a cognition, but insofar as it is ongoing it lacks an essential or determinate spatio-temporal location such that it could be properly cognized at any given moment.¹²⁰

§4.3. Historical Precedence for Model 3

The elements of this model inform the immediate post-Kantian German Idealist accounts of the self. These German Idealists, especially Fichte, identify the self *solely* with activity: that is to say,

processes and not events. And the language of ‘actor’ is helpful in specifying that the action is related to its author, even if, for instance, there is an intervening event that could be situated as the direct cause.

¹¹⁸ That we *cannot* cognize the self in inner sense is suggested, at least, by Kant’s claims that in inner sense there is no fixed object, e.g., A22/B37, A107.

¹¹⁹ See Kraus (2020, 3-6) for an overview of the debate about whether and how we can cognize and know the self in inner experience.

¹²⁰ On this view, we can cognize the individual states that compose the empirical self (the individual events or properties), but we cannot cognize it as a whole because we don’t yet have access to all of its determinations. Another upside of this model is that it is a metaphysical picture that is nevertheless compatible with self-constitution or self-formation views of self-knowledge, like those offered by Korsgaard (2009), Kraus (2020), or Frankfurt (1971). It complements these accounts that take it that what the self *is* is a process that we are involved in producing, and further responds to a worry that threatens these views, viz. what is the thing doing the producing, or the stringing together of representations, desires, and feelings, if it is not itself a self? Kant’s solution is straightforward: your self is both the actor and the activity of stringing together.

they drop the actor part of the equation.¹²¹ If we read the Actor/Activity model onto Kant, then the result is a clear and understandable progression of the metaphysics of the self from Early Modern rationalist conceptions of the self as a substance, to Kant's picture which takes the self to be both a real object (an actor) *and* an activity, to the German Idealist conception of the self as *just* an activity. This situates Kant squarely at the intersection of, e.g. Leibniz and Fichte's metaphysics of the self, which is exactly where you would expect to find him given the directions of influence and philosophical ideas.

More support from historical precedence can be found in Schopenhauer, who describes a quite similar account of two selves conceived of as, respectively, the *act* of the will and the *action* of the body.¹²² In putting forward this account, Schopenhauer takes himself to be interpreting Kant: more reason to think that we are on to something in arriving at a similar interpretation.

§4.4. How Model 3 Can Satisfy the Desiderata

Here is how this model can satisfy our desiderata.

Beginning with D1. Recall that the prescription of this desideratum was that it could explain how both selves can be properly called the referent of the 'I'. This model can satisfy this desideratum by joining Kant's claim that we attach the 'I', or the 'I think', to every mental representation with his claim that this 'attaching' is the *activity* of synthesizing or unifying a series of empirical mental representations—thoughts, beliefs, desires—under this common description as *my* representations.

¹²¹ See Fichte (2000, 4).

¹²² Schopenhauer (1969, Book 1 §21).

In *performing* this activity, we are *conscious*—as Kant says in the passage cited above—of ourselves as existing as real and as making possible this activity. The ‘I’, then, is doubled, but identified, because *in* being conscious of the activity that is the empirical self, we are conscious of some further self (some further I). The two awarenesses are inseparable. The experience is singular but the resultant awareness of selves is doubled.¹²³

Moving on to D2: how can this account accommodate the everyday experience of moral action, and how can it explain moral progress? I’ll explain this with a toy example or analogy of a baker baking a cake. On almost all accounts of the ontology of action, the baker (the actor) is distinct from the activity of baking, even while they are carrying it out. In so far as they are distinguished, they can be independently criticized. This is easy to see if you think of the product of the activity: you can say both that the baker did something wrong (added salt instead of sugar), and you can say that a cake tastes bad. Now of course, only in the former case are you *directly* ascribing responsibility. But the latter instance, too, can be understood as an ascription of responsibility insofar as you take the independent criticism of the activity to express a criticism of the acting agent. You are both evaluating the activity and—in virtue of evaluating that activity—holding the actor responsible for the wrong.

We can appropriately criticize and evaluate both the actor and activity, and we can evaluate them jointly. The on-going process of activity is crucial for the explanation of moral progress. Imagine, in our baking example, that you taste a batter and say, ‘that is disgusting, something has gone

¹²³ That they can be doubled in this way, I take it, is in part explained by the fact that one of the (viz., consciousness of the real self) is a kind of feeling as opposed to a kind of perception, as I argued in chapter 1.

wrong here'. Luckily, though, the baking process is on-going, and so in hearing the criticism of the activity, the baker then modifies their own behaviour and output. As a general principle, a criticism or evaluation of an on-going activity can occur partially independently from the actor, and the resultant change will modify the behaviour of the actor. In this particular case, criticism of the empirical agent should prompt a modification in the behaviour of the noumenal agent.

It is important to acknowledge that applying this analogy to the case of Kant's selves does mean that we need to conceive of the possibility of change produced in and by a timeless agent. But how to reconcile timelessness and change, or whether timelessness and immutability are equivalent, is certainly not solely a Kant problem: it is a problem faced by any theory of timeless agency, including several accounts of God's activity in the world. Kant, then, is certainly not without precedent in asserting that such activity and changeability is possible. And, if we can accept this, we can accommodate moral progress and reject any fatalistic determinism that comes with asserting that our character is fixed by an original timeless choice.

Finally, D3 comes fairly easily. On a plausible ontology of action, the actor and the activity are distinct entities, but the activity cannot exist without the actor: the activity *depends* on the actor. This, then, meshes both with Kant's claims that the persistence and individuation conditions of the intention and action of the actor can and do come apart from the persistence and individuation conditions of the activity, and with the mind-dependency of empirical objects.

§4.5. How Model 3 Avoids the Problems of Previous Models

I've argued that Model 3 is textually compelling and satisfies our three key desiderata. It also avoids the main problems identified for the previous two models. The empirical self is the

characteristic activity of the real self, and so it can be explained as distinct from the real self without being minimized or reduced entirely to the real self. In claiming that the real self is an actor, we needn't make any claims about its substantiality that violate epistemic humility, and agnosticism about the properties of this self is consistent with the pillars of the interpretation. Since the Actor/Activity model takes the two selves to be distinct, it does not encounter issues of clarifying how two such different things can stand in a composition relation. It does raise a new issue of how a nontemporal entity can have temporal agency, but this, at least, is not a problem that is unique to Kant. Finally, while in experiencing the activity of the self the two selves make up a singular *experience* as a human being, they do not so by virtue of composing a whole and so this model does not end up adding yet another self to the picture.

§4.6. Why Two Selves (Might Be) Better Than One

Taken all together, I argue we get a compelling and coherent way to make sense of all of Kant's claims about selves. This way comes with the following result. On my reading of Kant on the relation between selves, it is actually incorrect to say that the empirical and the real self are one and the same—it would be like describing the baking of the cake as the very same thing as the baker. This is contrasted to the substance/accident reading, whereby the empirical self is a property (or a set of properties) that inheres in the real self and so there is really only one self. It is contrasted, too, to the whole/part reading, which keeps it available that we can refer to *one* thing as the self, by which we mean that thing that is the composition of the empirical and the real self.

I began with a commonsense idea: that it seems strange to say that each of us has any more than exactly one self. I have spent the majority of the chapter explicating why my view is the most textually and (Kantian-) philosophically promising account available to us. However, it might

seem to many that this bullet—the loss of this basic intuition—is too big to bite. While the previous solutions may run into problems, those problems are worth the tradeoff of preserving the singularity of the self.

I disagree. I will conclude, then, with some thoughts on why we actually shouldn't be wedded to the synchronic singularity of the self, and the benefits of philosophical theories—like, as I read it, Kant's—that push on the intuitiveness of this claim about singularity.¹²⁴

What follows are speculative thoughts for extensions of the view, not considered conclusions. I hold, though, that these claims can be developed in a direction that is fruitful both for our understanding of Kant and for how Kant can be instructive for contemporary thinking about identity.

There are two different and distinct directions to go. The first is just intended to put a foot in the door of our intuitions about selves. I am not claiming that the following considerations map onto the relation between the empirical and the real self in Kant, only that they suggest that we are not as intuitively committed to the synchronic singularity of the self as it first seems. Here is this first direction. We often speak of having distinguishable selves: if I say to you that you are seeing my 'work self', or my 'professional self', you would understand what I mean. Further, it would be comprehensible—maybe even relatable—for me to say that my 'work self' comes apart from

¹²⁴ Such a view is distinct from David Lewis's four-dimensionalism about persons. Lewis (1983) argues that the self *perdures* and does not *endure*. I understand this as the view that there is a unique or singular self at a given moment and then another distinct, singular self at another moment, and we string those selves ("person-stages" or proto-subjects) together with a relation such that the relation that connects them constitutes a single self when considered across time. Even if you take the relation between selves to be weaker than that, four-dimensionalism about the self is a view that argues against the *diachronic* singularity of the self, but maintains the synchronic singularity. I am interested here in the possibility of a view that denies that there is synchronic singularity of the self. For worries about the ethical consequences of conceiving of the multiplicity of selves in this way, see Johnston (2017).

my ‘home self’. That this has greater significance than just a mere quirk of language is supported by a phenomenon that I label self-dissonance: that is, instances where you feel like you don’t belong in a certain setting or event. In many theories of the self, including parts of Kant’s, what it is to be your self is at least partially constituted by your experiences and the external relations you find yourself in. If that is right, these feelings of self-dissonance indicate not that your *one self* does not belong in the setting, but that there is *conflict* between your selves. Such conflict is only possible if we take these selves to be actually distinct—that is, if we hold that you have more than one self.¹²⁵

The second direction comes from considerations I brought up in chapter 1 about how we think we should answer the question ‘who are you’. As I argued there, to answer this question you might first appeal first to relational facts: that you are a sibling, or a citizen of a particular country, or work a particular job. But if pressed, you would not want to say that these facts constitute who you are (at least, that they do not exhaust your description) because they are contingent and each could have been otherwise. If *you* think *you* could work a different job, have been born in a different country, etc., then you must be tracking that there is some feature of *you* that is not

¹²⁵ In raising these questions and problems about reconciling and distinguishing between identities and senses of self, I am picking up on philosophical theories like those of, e.g., Lugones’ (2022) theory of “world-traveling” and Du Bois’ (1996) double consciousness. I take it, as well, that in holding that Kant’s metaphysics has the room to accommodate and even explain this kind of phenomena, my account can extend the self-constitution or self-formation accounts of self-knowledge (as discussed in n.120). These accounts present self-knowledge as a rational achievement in accordance with a normative (or even teleological) demand: the demand to conceive of one’s self as a unity. Such accounts provide many resources for making sense of self-knowledge. But the normative context means that aspects of our identity and self-awareness that don’t accord with this prescription—aspects that are contradictory or *dissonant*—cannot ultimately be included in our self-knowledge. They are anomalies or aberrations to be pruned out in achieving a unified self-knowledge. This, I hold, leaves out important complexities in our self-understanding: identity dissonances are, if not a feature, certainly not a bug. On my reading of Kant, these complexities are built into the metaphysics of the self. To gesture at a way to combine this metaphysical picture with the account of feeling I argued for in the first chapter, the real self is not a prescription, but instead something that is *felt*, a feeling consistent with knowing that an experience feels both attributable to you and at odds with who you take yourself to be.

exhausted by these relational facts. What I am trying to motivate here is that there could be great value in having both a conception of a self that is characterized by spatiotemporal experiences, *and* a self that is outside of space and time. This second self would account for what goes unexplained in self-descriptions that appeal only to relations or to generics, or in these counterfactual imaginings.

It is true that both of these sets of considerations are most often attended to or considered relevant in ethical or moral contexts. But I think this is precisely what makes Kant's theory so interesting: as I argued in chapter 1, he is trying to provide a metaphysical framework for identity that can accommodate these practical identity markers, like what we value and the nature of our character. Whether he fully succeeds is beyond my purview here: taking seriously the attempt is, I hold, worthwhile in itself. The result is that, far from the strange, mysterious, or just completely hopeless picture that many readers have painted of Kant's theory of two selves as two distinct entities, it actually offers an innovative way of actualizing the phenomenological experience that we have of our *selves*. Since a two-entity picture, as I have argued, represents our best accounting for the metaphysics of the self in Kant, the explanatory work and practical implications of this picture can be taken seriously and spelled out at greater length.

§5. Conclusion

I began with a puzzle about the self in Kant: does he think we have one self, or two selves? I have now argued that the answer is two selves. We can see that this should be our answer by way of responding to a more complex puzzle, namely, how can we reconcile three crucial desiderata that seem to pull in different directions? I have argued that this can be accomplished by conceiving of

the relation between the selves as the relation of actor to activity, where the real self is as an actor sustaining the process of the empirical self, modeled on the activity.

As I cautioned at the start, my proposal of the Actor/Activity Model is not intended as the final word on the subject. In particular, the way I have undertaken the investigation into solutions to the puzzles leaves open the door to establish further non-negotiable desiderata for this relation: the identification of such desiderata would prompt a reevaluation of the most appropriate model.

One important thing to recognize about the Actor/Activity model and a direction for further research is that it is difficult (if not impossible) to see how this particular relation could be generalized as an account of the relation that obtains between appearances and things-in-themselves. This, I suggest, is a feature of the account. It opens the door to thinking about Kant's metaphysics as one that prioritizes, or at least affords a special and unique status to, the self or the person. It is interesting, I contend, to explore the idea that there is (non-accidentally) a substantive difference between the way we understand the relation between the empirical and real self and the way we understand the relation between any standard appearance *x* and a thing in itself.¹²⁶ In future work, I aim to extend my arguments here to consider this question.

¹²⁶ This leaves space to hold that Kant thinks that the empirical and real self are numerically distinct but that there are some appearances and things in themselves that are not. That there is not consistency across cases is argued by Adams (1997), although he proffers almost the opposite conclusion (viz. that the self is identical but other appearances are not identical to things in themselves). My proposal is that, unlike what is suggested by Adams, this is a deliberate guiding principle of Kant's metaphysics, not an accidental or runoff consequence of his moral arguments. Likewise, the space is left open to hold that the self *is* an exemplar, again for considered reasons. In either the direction, the core of the argument is as follows. Since our epistemic position in the case of the human self is a privileged one, we are in a position to understand that relation in greater detail than in any other case, and so likewise we should (or should not) assume that the same kind of unity obtains between any related pair of appearances and things in themselves.

As it stands, the Actor/Activity model is responsive to the various and varying demands made on the metaphysics of the Kantian self. In being so responsive, it also opens up interesting ways to expand and apply Kant's account of the self, or, more properly, *selves*. In conclusion, then, two selves are *not* too many selves: sometimes, a second self is just what we need.

Conclusion

Self-Cognition, Self-Representation, and Self-Knowledge

1. Summary of the Argument

This dissertation has explored questions about the metaphysically real self in Kant's theory of the self in three different domains: cognitive, epistemic, and metaphysical. In each of these domains, I have argued both that the real self plays an integral role in Kant's arguments and accounts, and that we can understand the role played by the real self to be consistent with his core philosophical commitments, including his commitment to humility about our ability to cognize and know things as they are in themselves. In so doing, I have shown how we can properly—without contradiction or incoherence—include the real self in our best account of the self, self-consciousness, and personal identity in Kant.

These results are only the beginning of the project of properly understanding Kant's theory of the self. There are several questions that deserve but have not received comprehensive treatment here. In what follows, I briefly elaborate on two future projects that I have gestured at throughout the dissertation: one that extends my account of Kant's theory, and another that offers an application of Kant's theory to a contemporary question. I conclude by noting a set of open questions and further directions with which to explore the theory of the self I find in Kant.

2. Self-Cognition and Practical Cognition: Kant's Account

The first of these picks up on an issue that I raised and left open in chapter 2: the role of consciousness in cognition and how we should understand consciousness as a constituent element of the complex mental state of cognition. I focused there on the case of *real self-cognition* and argued that an episode of consciousness was an independent necessary condition on this

kind of self-cognition. A constitutive element of cognition of the real self is consciousness of a specific object, namely, the real self.

There is good reason to think that this case of cognition is an instance of the general kind of cognition. Particular instances of an epistemic state should adhere to the general description of the state, or else there would be no ability to point to a unified, definable kind. If this is the case, then insofar as consciousness of a self is an independent condition on self-cognition, so too should consciousness of an object be an independent condition on cognition, generally.

Scholars have noted that a referring consciousness is an integral constituent of Kant's main account of cognition. But they have tended, almost univocally, to lump this consciousness condition in with the intuition condition on cognition, or what is frequently referred to as the *givenness* condition: the presence to mind of an existing object in sensible intuition. The No Real Self-Cognition argument, though, establishes that we can have the consciousness of an object absent a corresponding sensible intuition. Insofar as there is a consciousness condition on cognition, then, it must be an *independent, additional* condition. Kantian cognition, as a general state, is predicated on: (1) an intuition, (2) concepts, and (3) consciousness.

Now, disambiguating between the intuition condition and the consciousness condition will be, in almost all theoretical cases, just conceptual bookkeeping. It is important bookkeeping, as it illustrates the special and unique epistemic status that 'consciousness' has in Kant. But the objects of theoretical cognition are ones where the givenness or intuition of the object goes hand-in-hand with the consciousness of it: I cognize my pen because there is a physical, sensible, existing object that I am conscious of representing.

When we move beyond the theoretical case, though, I argue that an independent consciousness condition is instrumental in explaining the difference between *theoretical* and *practical* cognition. Here is what I mean. In theoretical texts, Kant argues that cognition is restricted to those objects that appear in space and time and so we *cannot* have (theoretical) cognition of things-in-themselves. In practical writings, Kant allows that we *can* (practically) cognize (at least certain) things-in-themselves. While historically, interpretations of Kantian cognition have restricted themselves to explaining theoretical cognition, there has been recent and increasing interest in understanding the nature of practical cognition.¹²⁷ Nevertheless, there are still a number of open questions about practical cognition and about how practical cognition relates to theoretical cognition. Are these two entirely different states, or are these both species of a common genus, ‘cognition’? If the answer to the latter question is ‘yes’, what explains the difference between their possible content?

My basic proposal is that an isolated consciousness condition on cognition is precisely what we need in order to make sense of both the structural parity between theoretical and practical cognition (viz., they are both kinds of cognition) *and* of the significant difference in their possible content. Taking my cue from passages¹²⁸ where Kant describes practical cognition as unique

¹²⁷ In Kain (2010), Ameriks (2012), Abaci (2022), and Schafer (2023).

¹²⁸ For example: “Since in all precepts of pure practical reason, it is only a question of the decision of the will and not of the natural conditions (of practical ability) for achieving its purpose, it thereby happens that the practical concepts *a priori* in relation to a supreme principle of freedom immediately become cognitions not needing to wait upon intuitions in order to acquire meaning. This occurs for the noteworthy reason that they themselves produce the reality of that to which they refer (the disposition of the will)—an achievement which is in no way the business of theoretical concepts” (5:66). Existing interpretations attempt to make practical and theoretical cognition parallel by identifying an analogous role for intuition in the practical case, but it seems to me clear that Kant is claiming there is *no* role for intuition in practical cognition, analogous or otherwise.

because only in this case does cognition result *without any corresponding intuition*, I want to argue that in the theoretical case, cognition relies on empirical consciousness to satisfy the consciousness condition. Practical cognition, on the other hand, is the result of the consciousness criteria being satisfied by real consciousness. This is why there are limited objects available to practical cognition, and it can only occur in certain cases. But it is also the explanation for why practical cognition can be genuine cognition of supersensible objects, and why practical cognition is a genuine cognitive achievement (not a mere pragmatic assumption or an act of faith).

This account would tie together my claim that consciousness is a distinct and significant epistemic state in Kant with Kant's claims that recognition of the moral law allows us to extend our cognition further than the bounds of scientific and natural laws. I look forward to developing this argument and unpacking more elements of Kant's account of self-referring mental states.

3. Dissonance in Self-Representation: A Contemporary Application

The second thread I want to pick up on are my suggestions at the end of both chapters 1 and 3 that the cognitive and metaphysical implications of Kant's theory of the real self can be fruitfully adapted to explain and account for what it means to have multiple ways of representing oneself, or multiple *senses of self*. I've pointed to two different directions which we might take to arrive at the idea that we have a synchronic multiplicity of selves: that is, that at any moment, we have more than one self or at least more than one determinate identity. The first direction draws on considerations of what features or properties we could point to in order to answer the question 'who am I?'. The second direction draws on everyday language (e.g., "I have to be my *professional* self today"). I elaborated on this second direction in particular by pointing out that the idea that

an agent has multiple different senses of self or self-representations at one time is frequently mentioned in literature (it is the central contention, to name just one instance, of Virginia Woolf's *Orlando*), and it is the impetus for philosophical discussions of phenomenon like code-switching (Morton 2019, Velleman 2006), world-traveling (Lugones 2022), and double-consciousness (Du Bois 1996).

What animates these discussions are apparent hard cases: where senses of self are experienced simultaneously and in particular when they come into conflict, i.e., when an agent feels like a set of experiences or traits that she identifies with are incompatible or dissonant with another set with which she identifies. These feelings—I label the experience ‘self-dissonance’—indicate not that your *one self* does not belong in the setting, but that there is *conflict* between your selves. Such conflict is only possible if we take these selves to be actually distinct—that is, if we hold that you have more than one self. They suggest, at the very least, the need for a way to differentiate amongst these senses of self.

I hold that the account I have found in Kant throughout the dissertation can be employed to make precise what it is to have distinct senses of self by offering such an individuation criterion. I take it that in combining the metaphysics of the selves with the idea that our cognitive access to a facet of our identity is through the mode of feeling, Kant is well-situated to explain this phenomenon and to precisify the individuation of selves such that you can genuinely *feel* like you have distinct selves at one time. The experience of self-dissonance is an important facet of our self-knowledge, and so our account of self-knowledge should be responsive to these complexities. On my reading of Kant, these complexities are built into the metaphysics of the self. The ‘real self’ is something

that is *felt*, a feeling consistent with knowing that an experience is both attributable to you and at odds with who you take yourself to be. You don't need to buy into the metaphysics to get this value out of it, but if want a Kantian inspired account of self-knowledge that starting point is importantly revealing of this more complex picture.

I hope to develop this story at greater length as I continue to develop my project on Kant's theory of the self.

4. Open Questions and Future Directions

Both of these projects reflect a series of questions that were left open in the dissertation and deserve to be engaged with at length. I end here with a more rapid-fire offering of open questions and further directions with which to explore Kant's theory of the self. Some of these issues arose explicitly over the course of the dissertation, and others are in the background. The biggest of these is, of course, completing the full account of Kant's theory of the self: that is, making sense of the real self, the empirical self, their relation, and their joint function. This project arches over all of the issues I flag here: understanding each of them would represent a part of completing this full account.

(i) Taxonomizing Kinds of Consciousness and Apperception in Kant

I have been referring to the kind of consciousness that does not refer and structures all thinking and representation as *formal* consciousness. My reason for this was that, before we understand the concept itself, the technical terminology that surrounds it obfuscates rather than clarifies. Now that we have in hand clear understandings of the three kinds of consciousness in Kant's account of the mind, we are in position to return to this technical terminology and properly taxonomize it. In particular, we are now in a position to disambiguate between *pure*, *transcendental*, *mere*, and

empirical apperception and point out when the relevant kind is meant to be tracking real, formal, or empirical self-consciousness.

(ii) The Argument Structure of the Paralogisms

I have relied on the arguments in the Paralogisms of Pure Reason chapter to support my claim that Kant identifies a kind of consciousness that is formal only: merely the representation of a subject of thought, with no further properties. These arguments themselves, though, warrant further scrutiny. There is very little consensus on how to reconstruct the arguments in the Paralogisms such that they are evidence *of* a paralogism, that is, evidence of the fallacy of equivocation on the middle term. And when reconstructions of the arguments do preserve the fallacy of equivocation, they tend to prove too much: they would render Kant's own commitments in the Transcendental Deduction equally fallacious or unwarranted. The work I have done here in disambiguating kinds of consciousness in Kant can, I believe, shed new light on how we should best reconstruct and understand the syllogisms of the Paralogisms chapter.

(iii) The Self in Arguments for Idealism

I have here mentioned one way in which thinking about the metaphysics of the self can be useful for thinking about transcendental idealism, viz. in thinking about how and whether the relation that obtains between the real and the empirical selves generalizes over all instances of the relation between things in themselves and appearances. Another way in which the two projects are connected is in thinking through Kant's argument for the reality of the empirical world. In the B Preface, he offers a supplement to his refutation of psychological idealism—the “only possible proof”, as he puts it, of the reality of outer intuition (Bxxxix n)—that relies on several claims about our consciousness of our self and our existence. How should we understand these claims? How

does self-consciousness support Kant's proof of empirical reality, and how does this feature in his arguments for transcendental idealism?

(iv) The Mind and the Sublime

My arguments here have focused on how Kant's theory of the self is developed and supported across both theoretical texts—in particular, the *Critique of Pure Reason*—and practical texts (the *Groundwork for the Metaphysics of Morals*, the *Critique of Practical Reason*, and the *Metaphysics of Morals*). A natural extension of this theory is into the third and final Critical text, the *Critique of the Power of Judgement*. One clear place to apply the results of my arguments here is in thinking about what happens to the human mind when we encounter the sublime, especially the mathematically sublime. There, Kant writes that “even to be able to think the given infinite without contradiction requires a faculty in the human mind that is itself supersensible. For it is only by means of this and its idea of a noumenon, [...] that the infinite of the sensible world is completely comprehended in the pure intellectual estimation of magnitude under a concept” (CJ 5:255). What supersensible faculty of the human mind is responsible for thoughts of infinity? What kind of self-awareness do we have when we think in this way?

(v) Intellectual Feelings

Finally, in giving an account of real consciousness in Kant as analogous to a feeling-towards, I argued this was representative of an underexplored element of Kant's thinking, viz. his theory of intellectual feelings. Although the paradigmatic cases of *feeling* in Kant are sensible feelings, there are frequent instances—when discussing the moral law, necessity, and, of course, the metaphysically real self—that Kant appeals to a *feeling* with a non-sensible origin. A fruitful and interesting undertaking would be to offer a general and comprehensive account of intellectual

feelings in Kant, and how they cohere with sensible feelings. This would be worthwhile both for its own sake, and in understanding Kant's role as a predecessor to German Idealist thinkers. Several German Idealists and Post-Kantian philosophers—Fichte, Schelling, Schopenhauer, von Günderrode, and Kierkegaard—rely on *feeling* talk to articulate the way we relate to truths about the world and our own subjectivity. In spelling out Kant's account of feeling, we could better situate his influence on the concept of feeling in these later thinkers.

There is a great deal more to be said, written, and explored in each issue I have raised here. The dissertation, as it is, leaves these questions unanswered. I look forward to picking each of them up, and to advancing the project of a systematic understanding of the metaphysical, epistemic, cognitive, and ethical dimensions of Kant's theory of the self and self-referring states.

Bibliography

Works by Kant

All citations of Kant's works use the pagination of the standard Akademie edition (Akademie Ausgabe), cited by volume and page. Individual works are identified by the abbreviations listed below in brackets. I follow the practice of citing the *Critique of Pure Reason* by giving the page number for the first edition of 1781 (A) followed by the page number for the second edition of 1787 (B). Translations are from the editions listed below, with my occasional modifications or corrections made with reference to: Kant, Immanuel. *Kants Gesammelte Schriften*. Königlich Preussische Akademie der Wissenschaften. Vols 1-29. Berlin: De Gruyter, 1902. I mention the corrections only when they are relevant to the interpretation of Kant's arguments.

[A/B] *Critique of Pure Reason*. Translated and edited by Paul Guyer and Allen W. Wood. Cambridge: Cambridge University Press, 1998.

[Anth] *Anthropology from a Pragmatic Point of View*. Translated and edited by Robert B. Louden. Cambridge: Cambridge University Press, 2006.

[BL] *The Blomberg Logic*. In *Lectures on Logic*, translated and edited by J. Michael Young, 1-246. Cambridge: Cambridge University Press, 1992.

[CJ] *Critique of Judgement*. Translated and edited by Paul Guyer and Eric Matthews. Cambridge: Cambridge University Press, 2008.

[GM] *Groundwork for the Metaphysics of Morals*. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 2011.

[JL] *Jäsche Logic*. In *Lectures on Logic*, 521-640.

[KpV] *Critique of Practical Reason*. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 2015.

[ML] *Metaphysik L1*. In *Lectures on Metaphysics*, translated and edited by Karl Ameriks and Steve Naragon, 19-106. Cambridge: Cambridge University Press, 1997.

[MM] *The Metaphysics of Morals*. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 1991.

[Prol] *Prolegomena to Any Future Metaphysics That Will Be Able to Come Forward as Science*. Translated and edited by Gary Hatfield. Cambridge: Cambridge University Press, 2004.

Other Works

Abaci, Uygur. "Kant's Enigmatic Transition: Practical Cognition of the Supersensible." In *The Sensible and Intelligible Worlds*, edited by Nicholas Stang and Karl Schafer, 360-381. Oxford: Oxford University Press, 2022.

Adams, Robert Merrihew. "Things in Themselves." *Philosophy and Phenomenological Research* 57 (1997): 801-825.

Allais, Lucy. *Manifest Reality: Kant's Idealism and his Realism*. Oxford: Oxford University Press, 2015

Allison, Henry. *Kant's Transcendental Idealism*. New Haven: Yale University Press, 2004.

- — —. *Kant's Transcendental Deduction*. Oxford: Oxford University Press, 2015.
- — —. *Kant's Theory of Taste*. Cambridge: Cambridge University Press, 2001.
- — —. *Kant's Groundwork for the Metaphysics of Morals: A Commentary*. Oxford: Oxford University Press, 2011.
- Ameriks, Karl. *Kant's Theory of the Mind*. Oxford: Oxford University Press, 2000.
- — —. *Kant's Elliptical Path*. Oxford: Oxford University, 2012.
- Anscombe, G.E.M. *Intention*. Oxford: Basil Blackwell, 1957.
- Block, Ned and Ian Phillips. "Debate on Unconscious Perception." In *Current Controversies in Philosophy of Perception*, edited by Bence Nanay, 165-192. New York: Routledge Press, 2016.
- Boyle, Matthew. "Two Kinds of Self-Knowledge." *Philosophy and Phenomenological Research* 1 (2009): 133-164.
- Brook, Andrew. *Kant and the Mind*. Cambridge: Cambridge University Press, 1994.
- Burge, Tyler. "Two Kinds of Consciousness." In *Foundations of Mind: Philosophical Essays, Vol. 2*, edited by Tyler Burge, 383-391. Oxford: Oxford University Press, 1997.
- Charles, David. "Processes, Activities, and Actions." In *Process, Action, and Experience*, edited by Rowland Stout. Oxford: Oxford University Press, 2018.
- Chignell, Andrew. "Kant's One-World Phenomenalism: How the Moral Features Appear". In *The Sensible and Intelligible Worlds*, edited by Nicholas Stang and Karl Schafer, 337-359. Oxford: Oxford University Press, 2022.
- — —. "Can't Kant Cognize Himself? Or, a Problem for (Almost) Every Interpretation of Idealism." In *Kant and the Philosophy of Mind*, edited by Anil Gomes and Andrew Stephenson, 138-157. Oxford: Oxford University Press, 2017.
- — —. "Real Repugnance and Belief about Things-in-Themselves." In *Kant's Moral Metaphysics: God, Freedom, and Immortality*, edited by Benjamin J. Bruxvoort Lipscomb and James Krueger, 177-209. Berlin: De Gruyter, 2010.
- — —. "Kant's Concepts of Justification." *Nous* 41 (2007): 33-63.
- Choi, Yoon. "Spontaneity and Self-Consciousness in the *Groundwork* and the B-Critique." *Canadian Journal of Philosophy* 49 (2019): 936-955.
- Cohen, Alix. "A Kantian Account of Emotions as Feelings." *Mind* 129 (2020): 429-460.
- Della Rocca, Michael. "A rationalist manifesto: Spinoza and the principle of sufficient reason." *Philosophical Topics* 31: 75-93.
- Descartes, Réne. *Meditations on First Philosophy*. Translated and edited by John Cottingham. Cambridge: Cambridge University Press, 1996.
- Du Bois, W.E.B. *The Souls of Black Folk*. Project Gutenberg, 1996.
- Dyck, Corey. "Empirical Consciousness Explained: Self-Affection, (Self-)Consciousness, and Perception in the B Deduction." *Kantian Review* 11 (2006): 26-54.
- Emundts, Dina. "Die Paralogismen und die Widerlegung des Idealismus in Kants 'Kritik der reinen Verunft'." *Deutsche Zeitschrift für Philosophie* 295 (2006): 296-309.
- Evans, Gareth. *The Varieties of Reference*. Oxford: Clarendon Press, 1982.

- Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1971).
- Freitag, Wolfgang and Katharina Kraus. "An Expressivist Interpretation of Kant's I think." *Nous* 56 (2022): 110-132.
- Fichte, Johann Gottlieb. *The Foundations of Natural Right*. Translated by Michael Baur. Cambridge: Cambridge University Press, 2000.
- Ginsborg, Hannah. "The Appearance of Spontaneity: Kant on Judgement and Empirical Self-Knowledge." In *Self, World, and Art: Metaphysical Topics in Kant and Hegel*, edited by Dina Emundts, 119-144. Berlin: De Gruyter, 2013.
- — —. "Aesthetic Judging and the Intentionality of Pleasure." *Inquiry* 46 (2003): 164-181.
- Goldie, Peter. "Emotions, Feelings, and Intentionality." *Phenomenology and the Cognitive Sciences*. 1 (2002): 235-254.
- Grenberg, Jeanine. *Kant's Defense of Common Moral Experience*. Cambridge: Cambridge University Press, 2013.
- Grüne, Stefanie. "Kant on the Spontaneity of the Understanding." In *Self, World, and Art: Metaphysical Topics in Kant and Hegel*, edited by Dina Emundts, 145-176.
- Guyer, Paul. *Kant and the Claims of Knowledge*. New York: Cambridge University Press, 1987.
- — —. *Kant and the Claims of Taste*. Cambridge: Cambridge University Press, 1997.
- Heide, Dai. "A mereological argument for the non-spatiotemporality of things in themselves." *European Journal of Philosophy* 28 (2020): 48-76.
- Heimsoeth, Heinz. "Consciousness of Personality and the Thing in Itself." In *Kant: Disputed Questions*, edited by Moltke S. Gram, 237-277. Ohio: Ridgeway Publishing, 1984.
- Henrich, Dieter. *The Unity of Reason: Essays on Kant's Philosophy*. Cambridge: Harvard University Press, 1994.
- — —. "The Proof Structure of Kant's Transcendental Deduction." In *Kant on Pure Reason*, edited by Ralph C. Walker, 66-81. New York: Oxford University Press, 1982.
- Hogan, Desmond. "Handedness, Idealism, and Freedom." *Philosophical Review* 130 (2021): 385-449.
- — —. "How to Know Unknowable Things in Themselves." *Nous* 43 (2009): 49-63.
- Hume, David. *A Treatise on Human Nature*. Oxford: Clarendon Press, 1888.
- Indregard, Jonas Jervall. "Consciousness as Inner Sensation: Crusius and Kant." *Ergo* 5 (2018): 173-201.
- Ismael, J.T. *The Situated Self*. Oxford: Oxford University Press, 2006.
- Jacobi, J. Georg. "Über das Unternehmen des Criticismus, die Vernunft zu Verstande zu bringen, und der Philosophie überhaupt eine neue Absicht zu geben." In *Beyträge zur leichtern Uebersicht des Zustandes der Philosophie bey dem Anfange des 19. Jahrhunderts*, edited by C.L. Reinhold, 1-110. Hamburg.
- — —. "David Hume über den Glauben oder Idealismus und Realismus." Reprinted in *Jacobi, Werke Vol II*, 301-5. Leipzig: Gerhard Fleischer, 1815

- Jauernig, Anja. *The World According to Kant: Appearances and Things in Themselves*. Oxford: Oxford University Press, 2021.
- Johnston, Mark. "The Personite Problem: Should Practical Reason be Tabled?" *Nous* 51 (2017): 617-644.
- Kain, Patrick. "Practical Cognition, Intuition, and the Fact of Reason." In *Kant's Moral Metaphysics: God, Freedom, and Immortality*, edited by Benjamin J. Bruxvoort Lipscomb and James Krueger, 211-230. Berlin: De Gruyter, 2010.
- Kitcher, Patricia. *Kant's Thinker*. Oxford: Oxford University Press, 2011.
- — —. "Kant's Real Self." In *Self and Nature in Kant's Philosophy*, edited by Allen W. Wood, 73-101. Ithaca: Cornell University Press, 1984.
- — —. "Book Review: Kant on Self-Knowledge and Self-Formation by Katharina T. Kraus." *European Journal of Philosophy* 29 (2021): 679-682.
- Kleingeld, Pauline. "Moral consciousness and the fact of reason". In *Kant's Critique of Practical Reason: A Critical Guide*, edited by Andrews Reath and Jens Timmermann. Cambridge: Cambridge University Press, 2010.
- Korsgaard, Christine. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press, 2009.
- Kraus, Katharina. *Kant on Self-Knowledge and Self-Formation*. Cambridge: Cambridge University Press, 2020.
- Langton, Rae. *Kantian Humility: Our Ignorance of Things in Themselves*. Oxford: Oxford University Press, 1998.
- Lenczewska, Olga. "Expansion of Self-Consciousness in Kant's *Critique of Pure Reason*." *Kant-Studien* 110 (2019): 554-594.
- Lewis, David. "Survival and Identity." In *Philosophical Papers Vol. 1*. New York: Oxford University Press, 1983.
- Locke, John. *Essays on Human Understanding*. Edited by Peter Nidditch. Oxford: Oxford University Press, 1975.
- Longuenesse, Béatrice. *I, Me, Mine: Back to Kant, and Back Again*. Oxford: Oxford University Press, 2017.
- — —. "Kant on Consciousness and its Limits." *Revue de Métaphysique et de Morale* 117 (2023): 7-26.
- Lu-Adler, Huaping. *Kant, Race, and Racism: Views from Somewhere*. Oxford: Oxford University Press, 2023.
- Lugones, María. "Playfulness, "World"-Traveling, and Loving Perception." In *Feminist Philosophy of Mind*, edited by Keya Maitra and Jennifer McWeeny, 105-122. Oxford: Oxford University Press, 2022.
- Mariña, Jacqueline. "Transcendental Arguments for Personal Identity in Kant's Transcendental Deduction." *Philo* 14 (2011): 109-135.
- Marshall, Colin. "Kant's Metaphysics of the Self." *Philosophers' Imprint* 10 (2010): 1-42.

- — —. "Kant's One Self and the Appearance/Thing-in-itself Distinction." *Kant-Studien* 104 (2013): 421-441.
- McLear, Colin. "I Am the Original of All Objects": Apperception and the Substantial Subject." *Philosophers Imprint* 20 (2020): 1-38.
- Melnick, Arthur. *Kant's Theory of the Self*. New York: Routledge Press, 2009.
- Merritt, Melissa. *Kant on Reflection and Virtue*. Cambridge: Cambridge University Press, 2018.
- Morrison, John. "Spinoza on Mind, Body, and Numerical Identity." In *Oxford Studies in Philosophy of Mind Vol. 2*, edited by Uriah Kriegel, 293-336. Oxford: Oxford University Press, 2022.
- Morton, Jennifer. *Moving Up Without Losing Your Way: The Ethical Costs of Upward Mobility*. Princeton: Princeton University Press, 2019.
- Parfit, Derek. "Personal Identity." *The Philosophical Review* 80 (1971): 3-27.
- Pasnau, Robert. *Metaphysical Themes 1274-1671*. Oxford: Oxford University Press, 2011.
- Pereboom, Derk. "Kant on Intentionality." *Synthese* 77 (1988): 321-352.
- Piper, Adrian. *Rationality and the Structure of the Self*. Germany: APRA Foundation Berlin, 2013.
- Proops, Ian. *The Fiery Test of Critique*. Oxford: Oxford University Press, 2021.
- Rosefeldt, Tobias. "Kant's Self: Real Entity and Logical Identity." In *Strawson and Kant*, edited by Hans-Johann Glock, 141-154. Oxford: Oxford University Press, 2003.
- Ruben, David-Hillel. *The Metaphysics of Action: Trying, Doing, Causing*. Cham: Palgrave MacMillan, 2018.
- Schafer, Karl. "Practical Cognition and Knowledge of Things in Themselves." In *The Idea of Freedom: New Essays on the Kantian Idea of Freedom*, edited by Evan Tiffany and Dai Heide. Oxford: Oxford University Press, 2023.
- — —. "Kant's Conception of Cognition and Our Knowledge of Things in Themselves." In *The Sensible and Intelligible Worlds*, edited by Nicholas Stang and Karl Schafer, 247-278. Oxford: Oxford University Press, 2022.
- Schopenhauer, Arthur. *The World as Will and Representation, Volume I*. Translated by E. F. J. Payne. New York: Dover, 1969.
- Sethi, Janum. "Kant on Empirical Self-Consciousness." *Australasian Journal of Philosophy* (2021): 1-21.
- Smit, Houston. "Kant on Marks and the Immediacy of Intuition." *The Philosophical Review* 109 (2000): 235-266.
- — —. "Kant on Apriority and the Spontaneity of Cognition." In *Metaphysics and the Good: Themes from the Philosophy of Robert Merrihew Adams*, edited by Samuel Newlands and Larry Jorgensen, 188-251. Oxford: Oxford University Press, 2009.
- Strawson, P.F. *The Bounds of Sense*. New York: Routledge, 1966.
- Tolley, Clinton. "Kant on the place of cognition in the progression of our representations." *Synthese* 197 (2017): 3215-3244.

- Valaris, Markos. "Inner Sense, Self-Affection, and Temporal Consciousness in Kant's *Critique of Pure Reason*." *Philosophers' Imprint* 8 (2008): 1-18.
- Van Cleve, James. *Problems from Kant*. New York: Oxford University Press, 1999.
- Velleman, David. *Self to Self: Selected Essays*. Cambridge: Cambridge University Press, 2006.
- Walker, Ralph. "Kant on the Number of Worlds." *British Journal for the History of Philosophy* 18 (2010): 821-843.
- Watkins, Eric. *Kant and the Metaphysics of Causality*. Cambridge: Cambridge University Press, 2005.
- Watkins, Eric and Marcus Willaschek. "Kant's Account of Cognition." *Journal of the History of Philosophy* 55 (2017): 83-117.
- Ware, Owen. "Accessing the Moral Law through Feeling." *Kantian Review* 20 (2015): 301-311.
- Wittgenstein, Ludwig. *The Blue and Brown Books*. Oxford: Blackwell, 1958.
- Wood, Allen W. "Kant's Compatibilism." In *Self and Nature in Kant's Philosophy*, edited by Allen W. Wood, 73-101. Ithaca: Cornell University Press, 1984.
- Wuerth, Julian. *Kant on Mind, Action, and Ethics*. Oxford: Oxford University Press, 2014.