Not many philosophers would turn to Kant for a positive view about the metaphysics of the self (the referent of ‘I’). On the contrary, most of Kant’s interpreters read him as warning that any attempt to give a positive account of the self’s nature is doomed to failure, and as building his theories without metaphysical assumptions about the self. This broad interpretive approach, which I’ll call the “anti-metaphysical interpretation,” often sees Kant’s project as anticipating Wittgenstein’s claims that the self or subject “doesn’t belong to the world, but is a limit of the world.”

In some form or other, the anti-metaphysical interpretation is present in all major discussions of Kant’s views on the self. In a 1993 survey of the literature, Günter Zöller stated that “[d]ifferences of method and philosophical approach aside, the interpretations … show a remarkable agreement in their understanding of Kant’s thinking self as a form or structure that eludes any attempt at reification.” The consensus now is similar — for instance, in the most recent book on the topic Arthur Melnick argues that in Kant’s theoretical philosophy the self should be understood as something like an activity precisely because this avoids construing it as an entity of any sort.

In what follows, I argue that not only is the anti-metaphysical interpretation mistaken, but that Kant offers us a subtle, plausible metaphysical account of the self that has no direct analogue in the contemporary literature. In a nutshell, the view is this: selves are individuated...
by experiences, so that whatever entities are jointly responsible for unifying sensations into a single experience thereby constitute a single self.

I proceed as follows. In Section 1, I clarify what in Kant’s theory I’m calling the ‘self’, I distinguish two senses of ‘metaphysics’, and I argue that the anti-metaphysical interpretation should not be the default view. Section 2 presents what I think are the two best arguments for the anti-metaphysical interpretation, shows how neither argument is decisive, and then argues directly for a metaphysical interpretation. Section 3 considers the most obvious candidates for a Kantian metaphysics of the self, and concludes that none of them is satisfactory. Section 4 presents my own interpretation, appealing to both intuitive and textual support. I conclude by briefly noting some of the broader philosophical virtues of this view.

1. Preliminaries

1.1 The Self, the I, the Mind and the Thinking Subject

I’ve described my topic as Kant’s metaphysics of the self or of the referent of ‘I’, but in the relevant passages from Kant and in the related secondary literature there is also talk of the metaphysics of the I (das Ich), the mind (das Gemüt), the thinking subject (das denkende Subjekt), and the soul (die Seele). In addition, Kant uses both the substantive form of ‘self’ (‘Selbst’) and the simple reflexive form (‘selbst’).

Kant never gives a clear statement of the relations between most of these terms. But on the present topic, he uses them all in effectively the same way.5 Two points are in order: (1) Kant recognizes a broad use of ‘I’ that refers to the “whole man,” involving body and soul.6 But Kant’s use of ‘I’ in the Critique of Pure Reason is meant in a more strict sense: “I, as thinking, am an object of inner sense, and am called ‘soul’” (A243/B400).7 In what follows, I’ll be concerned with the notion of ‘self’ that is associated with this narrower sense of ‘I’. (2) As will become clear, Kant holds that having a self requires more than merely possessing representations — it requires having certain faculties. But though he thinks animals lack these faculties, and so lack selves, Kant ascribes representations to them, so he must hold that animals have a mind of some sort.8 My topic, then, will be the sort of mind, referred to by ‘I’, which we have and animals lack.

1.2 metaphysics and Metaphysics

In talking about the “anti-metaphysical” interpretation, I intend to use the word ‘metaphysics’ in the sense prevalent among contemporary analytic philosophers, according to which a metaphysics of the self would be some account of how selves are individuated, what sort of entities they are, and how they fit into our broadest conception of reality. It’s not clear whether this contemporary sense of ‘metaphysics’ can be given a precise definition, but our intuitive understanding of it will be enough for what follows.

Kant himself states that “this name [‘metaphysics’] can also be given to all pure philosophy including this critique” (A841/B869), but he also uses it to describe a specific dogmatic enterprise aimed at establishing the reality of God, freedom, and immortality (cf. B7). This latter

5. Cf. the use of ‘I’, ‘mind’, ‘subject’, and ‘soul’ at B68-69, and the introduction to the Paralogisms: “I, as thinking, am an object of inner sense, and am called ‘soul’ … Accordingly, the expression ‘I’, as a thinking being, already signifies the object of a psychology that could be called the rational doctrine of the soul” (A243/B400). ‘Subject’ for Kant has a variety of senses (see Rosefeldt 2000, 33–35), but in most of the passages I consider, it is equivalent to the other terms.


7. Unless otherwise noted, all English translations are from Paul Guyer and Allen Wood’s translation of the Critique of Pure Reason. As is standard, I refer to passages from the Critique by indicating the page number of the first and second editions (e.g., A277/B333, and passages from other parts of Kant’s writings by indicating the title with the volume and page number of the Berlin Academy edition (e.g., Prolegomena 4:348).

sense of ‘metaphysics’ describes the discipline that Kant criticizes in the Transcendental Dialectic.

These distinctions can help us avoid a simple mistake: even if Kant is clearly opposed to the ambitious, dogmatic sort of metaphysics, that alone doesn’t show that he is opposed to metaphysics in the contemporary sense. To make this distinction clear, in the rest of this paper I use ‘Metaphysics’ to refer to the discipline Kant explicitly opposed, and ‘metaphysics’ to refer to the broader enterprise we today think of as metaphysics (leaving aside the question of how Kant’s own, more inclusive sense of ‘metaphysics’ lines up with ours). Kant is clearly an anti-Metaphysician about the self, but that doesn’t entail he is an anti-metaphysician about the self. This distinction will become clearer below, when I survey particular proposals concerning the self’s nature.

1.3 Why the anti-metaphysical approach needs support

In the next section, I consider some arguments that an advocate of the anti-metaphysical interpretation might advance. Before doing so, I want to make a few remarks about why an anti-metaphysical interpretation of Kant’s view of the self should not, for all its popularity, be the default view. Every one of Kant’s major predecessors, even those who were skeptical about Metaphysics, offered a clear metaphysical view of the self, so a complete rejection of positive metaphysics would be a dramatic enough shift that we would expect Kant to have announced it. Since he didn’t, the onus is on the side of the advocates of the anti-metaphysical interpretation.

Now, one reason why Kant is commonly interpreted in the anti-metaphysical way is that his theoretical philosophy is often read as a response to a strong form of skepticism. If Kant’s arguments were aimed at a skeptic who admits nothing other than her own current representations, then, given the potency of Humean worries about an underlying self, the most charitable way to interpret Kant would be as making no metaphysical assumptions about the self at all. This way of reading Kant has some textual support, and has been adopted by many of his interpreters, but it is nevertheless not obviously right. With an eye towards Kant’s insistence that it never came into his head to doubt the existence of outer objects (Prolegomena 4:293), Karl Ameriks has argued that Kant is best read as assuming a fairly robust conception of experience and reality, and prominent recent interpreters (perhaps mostly clearly, Rae Langton and Desmond Hogan) have followed Ameriks in attributing to Kant a richer background metaphysical picture than would make sense for someone who aims to refute skepticism on its own terms.

The interpretive issues here are difficult, but it is clear that it should not go without saying that Kant’s project required him to avoid metaphysical claims of all forms. To decide whether Kant has a metaphysical view of the self, we must consider more specific arguments for the anti-metaphysical interpretation.

2. For and against the anti-metaphysical interpretation

This section argues that we should reject the anti-metaphysical interpretation. After surveying and rejecting what I see as the two best

9. Hume, for instance, was plainly skeptical of Metaphysics, and yet (even though he appeared to have had doubts about it later) offered a metaphysical account of the mind as a bundle or heap of perceptions.

10. For instance, this approach is assumed in Paul Guyer’s criticisms of several arguments he finds in the Deduction, namely, that they would be question-begging against a skeptic (e.g., Guyer 1987, 86). It is also a chief motivation for attributing ‘transcendental arguments’ to Kant (see Walker, 2006), and something like it is the motivation for Patricia Kitcher’s interpretation of Kant on the self (see Kitcher 1994, Chapter 4). Michael Forster has recently argued against reading Kant as responding to Cartesian “veil of perception” skepticism (in Forster 2008), though he still goes to lengths to find some sort of skeptical motivation for the Kantian project.

11. Ameriks 1978 (Ameriks’ argument is significantly strengthened when paired with Engstrom 1994). Ameriks gives an overview of his approach in the introduction to Ameriks 2003. With respect to the metaphysics of the self, however, Ameriks seems relatively cautious, saying that we should attribute to Kant only those metaphysical claims about the self that are “forced on us by more general aspects of Kantian ontology” — in particular, the property of immateriality (Ameriks 2000, 308).

arguments for the anti-metaphysical interpretation, I present two arguments for adopting some sort of metaphysical interpretation. All this will leave open the possibility that Kant’s view is simply inconsistent (a view taken by many of Kant’s readers);10 it is the task of Section 4 to show that there is an account of the self we can attribute to Kant that is consistent with the important parts of his system.

The arguments for the anti-metaphysical interpretation I consider are based, first, on Kant’s claims about the limits of cognition and objective significance, and second, on Kant’s rejection of rational psychology. Though these seem to be the best arguments for the anti-metaphysical interpretation, neither is ultimately successful.

2.1 The first argument: Kant’s epistemological and representational strictures
Kant’s discussions of the self show he thinks that, if there were any such thing, it would have to be something outside of appearances, in the realm of things in themselves. Since that is a realm we could only cognize a priori, a metaphysics of the self would have to be an a priori metaphysics of a thing in itself. But one of Kant’s central doctrines is that we have no cognition of things in themselves (e.g., A42/B59ff.), and he explicitly carries this doctrine over to the case of the self: “I ... have no cognition of myself as I am, but only as I appear to myself” (B158).14 Since a metaphysics of the self would have to involve some sort of claim concerning the nature of the self in itself, it would seem that Kant has ruled out the possibility of any such doctrine’s being legitimate.15

13. This is a conclusion that has been drawn by many with respect to the general relation between Kant’s talk of things outside of representations and his metaphysical strictures. Perhaps the earliest example was Jacobi’s 1787 assertion that without the thing in itself, he couldn’t enter the critical system, and yet with it, he could not remain there (Jacobi 1788, vol. 2, 304).
15. As evidence of the influence of this argument: even Robert Howell, a reader very sympathetic to Kant’s views on the self, states that we should “separate Kant’s specific, core claims about the I think from the further views he reaches once his idealism is adopted” (Howell 2001, 119).

A more radical version of this argument can be made on the basis of Kant’s claims concerning the limits of the use of the categories. On this version of the argument, not only must we lack knowledge of the metaphysics of the self, but we are unable to even represent facts about the self as it is in itself. In his discussion of the distinction between phenomena and noumena, Kant explicitly states that, without the possibility of a corresponding intuition, a concept “has no sense, and is entirely empty of content” (A239/B298), that without empirical intuitions concepts “have no objective validity at all, but are rather a mere play” (A239/B298), and that without conditions of sensibility, “all significance, i.e., relation to the object disappears” (A241/B300).16 But since intuition and sensibility only yield cognition of appearances (cf. A42/B59), an a priori metaphysics of the self would have to involve a use of concepts without intuition or sensibility. Given that, Kant seems to be committed to claiming that such a doctrine would be not just illegitimate, but nonsense.

2.2 Why the first argument fails
Let’s start with the less radical version of the above argument, which focuses on Kant’s claims about the limits of our cognition. This version works best on the assumption that Kant’s key claim (namely, that we have no cognition of our selves as we are in ourselves) means that we have no knowledge of any facts about our selves outside of how we appear. There are reasons for doubting this assumption, however, and several recent commentators have argued for more nuanced understandings of Kant’s claims about cognition. It isn’t essential for my purposes to consider these proposals in detail.17 Instead I’ll just indicate three general issues that undermine the crucial assumption.

16. Such passages are the motivation for verificationist readings of Kant, the most famous being Strawson’s attribution to him of a verificationist “principle of significance” (Strawson 1966, 16–17).
17. Schafer (unpublished), Adams 1997, and Smit 1999 argue that a close understanding of the notion of Erkenntnis is needed when assessing Kant’s epistemic claims, though Adams and Smit take a somewhat different approach to the notion than the one mentioned here (focusing on Kant’s talk of real vs.
The first ground for doubting whether Kant means to say that we know no facts about ourselves outside of how we appear is just that Kant himself makes claims about how we are outside of how we appear. For instance, he asserts that our sensations are the result of objects affecting our minds (cf. A19/B33) and that the combination of representations can only be executed by the subject itself (cf. B130). Even Kant’s claim that we do not know ourselves as we are in ourselves presupposes that we have a nature in ourselves, and this certainly does not seem to be merely a claim about how we appear. This implication is clear in his statement “[this] soul I cognize only as an object of internal sense through the appearances constituting an inner state, and whose being as it is in itself, which underlies these appearances, is unknown to me” (Prolegomena 4:336). I discuss more examples of such claims in Section 2.6, below.

The second ground for doubting the assumption comes from Kant’s own claims about cognition (Erkenntnis), the term he typically uses in expressing these limitations. Kant is quite clear that ‘cognition’ is a technical term. In the Deduction, he states that cognition consists “in the determinate relation of given representations to an object” (B137, see also B146–74) and in the Introduction to the B edition he describes the production of cognition as involving the comparison of different sensory representations. A similar point is made in the Jäsche Logic at AA 9:65. Whether or not one takes such passages as demonstrating what Kant means by ‘cognition’, they do give us reasons to question

---

*logical possibility.* In recent work, Desmond Hogan explores a very different approach to this issue, one which, if ultimately defensible and generalizable, would also provide a response to the above argument (Hogan 2009a, Hogan 2009b, Hogan 2009c). Something similar is true of Langton 1998.

18. Bt. In other places, Kant uses ‘Erkenntnis’ more liberally. Cf. A320/B376–77, where an intuition or concept alone is counted as a cognition, and the places in the logic lectures where any representation of something is counted as cognition (e.g., Logic Lectures 24:132, 24:845, and Jäsche Logic 9:64). But this broader understanding of the notion makes it impossible to make sense of Kant’s central claims concerning cognition in the Critique (such as the claim that it requires both intuitions and concepts at A51/B75).


---

20. Elsewhere, Kant puts his point by saying that “we do not and cannot have the least acquaintance [Kenntnis]” with the real subject of our thinking (A350). This term even more clearly calls out for interpretive care.

21. The verb ‘erkennen’, like its root ‘kennen’, takes objects as its objects, not propositions. The significance of this general point about Kant’s claims is being explored by Karl Schafer in unpublished work. I’m indebted to Schafer for helping me appreciate the need for care in reading Kant’s claims about Erkenntnis.

22. This familiar point was even recognized by Strawson (e.g., Strawson 1966, 240–41), who concluded that the project didn’t ultimately make sense.
Kant’s revisions to the Phenomena and Noumena chapter in the second edition of the *Critique* were intended to clarify this exact point.

On (1): immediately after saying that the categories have “relation to an object” only via the senses, Kant clarifies what he means by denying “validity” (*Gültigkeit*) to extra-sensory uses of the categories: “without the condition of sensible intuition … the categories have no relation at all to any *determinate* object, thus they cannot define one, and consequently they do not have in themselves any validity as concepts” (A246, my emphasis). Throughout the chapter, in fact, Kant often puts his point about the validity of concepts in terms of their relations to *determinate* objects, and as we should now expect, in terms of cognition. If Kant clarifies his claims about the limits of objective significance in the same way that he clarifies his claims about the limits of cognition, then his claims about the former are probably no stronger than his claims about the latter. Since we’ve seen that there is reason for taking his claims about cognition in a relatively weak sense, there is also reason for taking his claims about the limits of meaning and objective significance in a similarly weak sense. Once we do that, the argument for the anti-metaphysical interpretation loses its force.

On (2): there is also evidence that Kant realized he had been misleading in the A edition’s Phenomena and Noumena chapter. In at least 23. For instance: “though a pure category, in which abstraction is made from any condition of sensible intuitions … no object is determined, rather only the thought of an object in general is expressed in accordance with different *modi*” (A247/B304); “the other things, to which sensibility does not reach, are called noumena just in order to indicate that those cognitions cannot extend their domain to everything that the understanding thinks” (A254/B310); and “If we separate [intuitions and concepts], then we have … representations that we cannot relate to any determinate object” (A258/B314).

24. While Karl Ameriks mistakenly suggests the problematic passages are confined to the B edition, he does gesture at both this response, and (following Kemp Smith) the point of the next paragraph, when he notes that Kant had a ‘tendency to say something is wholly ‘without meaning’, or ‘significance’, when what he means is merely that it does not provide knowledge’ (Ameriks 2000, xxxv). Hogan 2009a charges Ameriks with adapting a “non-literal” reading of Kant’s claims, but this charge rests on his inherited assumption that Kant’s claims about *Erkenntnis* should be “literally” understood as claims about knowledge.

25. At the beginning of the chapter, at A237/B296, at A246/B303, twice at A247/B304, at A251, at A253 and at A259. This is detailed in Erdmann 1881 and in the Guyer/Wood translation.

26. This connection between determinate representation and meaning is also quite clear with regard to the self in the *Metaphysical Foundations of Natural Science*: “The I … itself merely a thought, designates, as a mere prefix, a thing of undetermined meaning — namely, the subject of all predicates — without any condition at all that would distinguish this representation of the subject from that of a something in general” (Metaphysical Foundations 4:542).
I conclude, then, that the two most promising arguments for the anti-metaphysical interpretation are inadequate. Just as importantly, though, there are strong, straightforward grounds for accepting a metaphysical interpretation of Kant’s views of the self. Both Kant’s overall project and the specific claims he makes appear to commit him to having some positive account of the self. I argue for each point in turn.

2.5 The metaphysical commitments of Kant’s Copernican revolution
Various interpreters have argued that Kant’s whole “Copernican” approach to philosophy must stem from a conception of a determinate mind with faculties.\(^\text{27}\) For instance, consider one of the more dramatic claims made in the A edition: “Thus we ourselves bring into the appearances that order and regularity in them that we call nature, and moreover we would not be able to find it there if we, or the nature of our mind, had not originally put it there” (A125). The project makes sense only if we understand the mind as being a genuine entity with a nature of its own that is more basic than the world of appearances, and where at least some aspects of that nature are known a priori.\(^\text{28}\) This point is in fact captured by Kant’s descriptions of “logical” features of the self or self-consciousness (e.g., A350, B413), for “transcendental logic” is described as the discipline that sets out to “isolate the understanding … and elevate from our cognition merely the part of our thought that has its origin solely in the understanding” (A62/B87). The self or mind is the object of investigation in the Critique — so Kant surely must have had some positive conception of what he was investigating.


\(^{28}\) In contrast to the view I favor, Waxman thinks that there is no good solution to the interpretive problem, and so concludes that “one is obliged to conclude that Kant’s fundamental descriptive categories — ‘representation’ (entailing a ‘represented’) and ‘thing in itself’ — lack any warrant” (Waxman 1990, 290). Patricia Kitcher considers a similar worry and concludes that it would be best to abandon Kant’s view of the self as noumenal (Kitcher 1984, 127).
2.6 Commitments to non-trivial facts about the self’s identity

The metaphysically committing claims concerning the identity of the self appear primarily in the Transcendental Deduction. Many interpreters of the Deduction, driven by the anti-metaphysical interpretation, read this stretch of the Critique in a non-metaphysical way,\(^{29}\) but I believe that this can be done only by seriously warping the text.

There is little question that one of the central notions in the Deduction is the “unity of apperception,”\(^{30}\) which Kant (somewhat more intelligibly) also describes as the “identity of self-consciousness.”\(^{31}\) Among other things, Kant states that this unity is “necessary” in some sense (e.g., A107, B142), that it is essentially connected to our fundamental concept of an object (e.g., A109, B139), that concept-guided synthesis and/or judgments are the means for bringing cognitions into this unity (e.g., A111-12, B141), and that we have some a priori conception of this unity as applied to the manifold of sensible intuition (e.g., A118, B150).

At least some of these claims are meant to be non-trivial, for they are pivotal steps in the Deduction, and Kant’s aim there is non-trivial (justifying our use of the categories). The important point is this: unless Kant is either invoking a new philosophical primitive (and there is no clear evidence that he is) or somehow reifying self-consciousness, there is no plausible way to understand what it is for there to be “identity of self-consciousness” over different representations without understanding it as involving a single conscious self that is the same in the different representations. We should not assume that we have an independent conception of what the “unity of apperception” amounts to. In the B Deduction Kant even introduces the notion by reference to a single self or subject: “all manifold of intuition has a necessary relation to the I think in the same subject in which this manifold is to be encountered” (B132, my emphasis). In the A Deduction he cashes out the notion in terms of “the identity of oneself” (A108).\(^{32}\)

Dieter Henrich, in his widely influential interpretation of the Deduction, resists this, asserting that “the identity of self-consciousness is not sufficient to enable one to infer the identity of a real thing or even of an object which is knowable with respect to its real nature. Numerical identity belongs to self-consciousness as such.”\(^{33}\) Similarly, Henry Allison, when considering the Paralogisms chapter of the Dialectic, holds that the illusion Kant there reveals is “to consider the unity of apperception as if it were the unity of a thing.”\(^{34}\) Both interpreters assume that we have a grasp on what it is for there to be identity or unity of self-consciousness without thinking of there being an identical or unified self. But reflection will show that we have no such grasp; if there’s really no thing at all that’s identical or unified over different states, then there just can’t be identity or unity of consciousness in those states.

Importantly, saying that there must be something that’s identical doesn’t itself presuppose any of the views Kant attacks in the Paralogisms chapter, namely, that we know that thing is a substance (First Paralogism), or simple (Second Paralogism), or distinct from material

\(^{29}\) See also A103, where apperception is clarified relative to facts about changes in a mind’s representational state.

\(^{30}\) Henrich 1994, 264. Henrich explicitly points the Paralogisms as showing this. Even the more metaphysically-minded Karl Ameriks says that the passages from the Deduction “have to be read quite carefully lest Kant be taken to be making precisely the paralogistic fallacy he clearly criticizes elsewhere in the Critique,” and goes on to claim that “the ‘identical’ item here need not be a thing but can be merely a formal condition or its representation” (Ameriks 2000, 140).

\(^{31}\) Allison 2004, 340. Tobias Rosefeldt argues that the Deduction passages only concern “the logical identity of the I,” where this needs to be distinguished from the identity of a thing in different states (Rosefeldt 2000, 128).
things (Fourth Paralogism). To be sure, the Third Paralogism is concerned with identity, but the issue is our identity as persons, and Kant talks freely there of our identity as selves (in the A edition at A361–366) or as subjects (in the B edition at B408).\footnote{Kant’s notion of ‘person’ (Person) then, is more loaded than Locke’s, since identity of self doesn’t entail identity of person for Kant. This point has been overlooked by previous commentators (most influentially, in Strawson 1966, Ameriks 2000, and Sellars 1970). I defend this interpretation of the Third Paralogism at length in my “The Identity of the Self in Kant’s Third Paralogism” (manuscript). Note that, with the exception of some very Kant-inspired work, all contemporary discussions of unity of consciousness assume that conscious states can be unified only if those states are states of the same mind.}

If we clarify our reading of the Deduction in this way, then we can recognize other passages that also make clear metaphysical claims about the self. Here are two such passages (others will be quoted below in Section 4.2). The italics are mine:

\begin{quote}
Even this representation — that all these appearances and thus all objects with which we can occupy ourselves are all in me, i.e., determinations of my identical self — expresses a thoroughgoing unity of them in one and the same apperception as necessary (A129).
\end{quote}

I am therefore conscious of the identical self in regard to the manifold of the representations that are given to me in an intuition because I call them all together my representations (B135).

Since the Deduction is supposed to be a priori, Kant is saying here that we have an a priori grasp on what it takes for us to have a unified, identical self with respect to some set of representations. Below, I say more about what that grasp might consist in. For now, the point is that there are commitments to non-trivial metaphysical claims about the self in the very chapter that Kant asserts is the most important “for getting to the bottom of that faculty we call the understanding” (Axxvi).\footnote{Wuerth 2006, 506–508 discusses how in the 1770s Kant insisted on these distinctions specifically with regard to the self. In particular, Kant criticizes the Wolffians for blurring the distinction between substances and powers and for construing the soul as a power (cf. Metaphysics L1, 28:261).}

Now, the arguments just given don’t tell us what positive account of the self Kant might have had. Part of the motivation for the anti-metaphysical interpretation has been that it is unclear what metaphysical account of the self could consistently fit into Kant’s system. Before I present my proposal, the next section will illustrate the relevant difficulties by considering the obvious candidates for the account.

3. The metaphysics of the self: some possibilities

Kant uses the same basic ontological categories as his predecessors in discussing the nature of the self, and in discussing the limits of cognition and of the significance of the categories.\footnote{This isn’t to say that Kant thinks that all applications of those categories are justified. Cf. Metaphysics Mroongovius 29:849–851. Kant’s critique of Leibniz’s monadology, for instance, is focused on Leibniz’s claims concerning the nature of substances, and the only objection to the application of the concept ‘substance’ in the passage concerns getting determinate knowledge of an object (recall Section 2.2). In this, I agree with Ameriks 1992 and Chapter 3 of Langton 1998.} On that ontology, the world consists of substances and their accidents. An especially important class of accidents is that of powers or forces, whereby substances influence each others’ accidents. Representations are a species of accident, as are relations.\footnote{I am therefore conscious of the identical self in regard to the manifold of the representations that are given to me in an intuition because I call them all together my representations (B135).} In what follows, I use ‘entity’ as a blanket term for both substances and accidents.

Using these terms, we can describe the most obvious candidates for an a priori metaphysics of the self. Following Descartes and Leibniz, Kant could have held that the self was a substance. Following Hume, he could have held that it was a bundle of representations. Two other candidates appear in the literature: the self as an activity or act, and the self as a form or structure. I consider each in turn. In the course of
3.1 The self as a substance?

Leibniz and Descartes hold that the self (that in me which thinks) is a mental substance. Bound up with this conclusion is the claim that the self is immortal, for part of what distinguishes genuine substances from accidents is their strong ontological independence.

Now, in his pre-Critical work, Kant does talk of the self’s being a substance. Moreover, much of Kant’s way of talking about the self in the Critique suggests continuity with the earlier rationalists’ views. Like Descartes, he consistently favors the first-person singular in talking about the nature of the self. He also seems to think that there is something special about entities capable of using the first-person. In the Anthropology, he says that the fact that a human can use ‘I’ ‘raises him infinitely above all other living beings on earth. Because of this he is a person, and by virtue of the unity of the consciousness through all changes that happen to him, one and the same person’ (Anthropology 7:127). Part of the traditional explanation for the specialness of humans was that they possessed immaterial, simple, immortal, intellectual souls.

At the same time, it is clear that in the Critique Kant does not accept Descartes’ and Leibniz’s view of the self’s metaphysical nature. Kant unambiguously attacks the warrant to any such claim in the First Paralogism (A348–351, B411–12). His precise target is the claim that we know a priori that the self is a substance, especially insofar as this carries the implication of immortality. This isn’t to rule out that we might in fact be one or more substances; the point is just that we have no a priori warrant for regarding ourselves as such. Indeed, he employs doing so, I identify the main desiderata for a plausible interpretation of Kant’s views on the self.

3.2 The self as a bundle of representations?

Hume famously claimed that the mind is nothing but a “bundle or collection of different perceptions.” Patricia Kitcher, in arguing that the Transcendental Deductions are a response to Hume’s challenge to the notion of personal identity, has described the Kantian I or self as “a contentually interconnected system of mental states.”

The possibility of the self’s ultimately being a collection or series of substances to illustrate the mistakes of rational psychology in the Second and Third Paralogisms (A353 and A363–64n., respectively).

But Kant ultimately thinks we cannot even assume that we are constituted by one or more substances, for he explicitly insists that our a priori awareness of ourselves does not determine “whether I could exist and be thought of only as subject and not as predicate of another thing” (B419). It is likely that Kant has in mind Spinoza’s view that the mind is an accident of God, the one substance. This means that, at least on his more considered view, Kant would deny that we have any a priori basis for affirming or denying that we are a simple entity, and that we have any such basis for thinking of our selves as being substances rather than accidents. If Kant has some a priori conception of the nature of the self, it would have to remain neutral on both issues.

39. For a recent survey, see Wuerth 2006.
40. This is setting aside justifications stemming from moral considerations, such as our need for supposing ourselves to be immortal in order to think of our endless progress towards moral perfection (see Critique of Practical Reason 5:122).
42. Hume 1778, I.vi.VI.
44. Kitcher 1984, 117 (see also Kitcher 1990, 122, and Kitcher 1998, 79). In her earlier work, Kitcher was clear that this was at best a partial reflection of Kant’s own view, since it made the self ultimately phenomenal (see Kitcher 1984, 121 and 134, and Kitcher 1990, 121, where the account is described as “deliberately revisionist”). She holds, however, that Kant’s main motivations for making the self noumenal are external to his theoretical philosophy (e.g., Kitcher 1990, 139), and the above arguments show this is a mistake. With a similar anti-metaphysical approach, Andrew Brook has claimed that Kant “did not just think of the mind as having a system of representations; he also thought of it as being a representation, namely, the global representation within which many of the usual denizens of a system of representations are all contained”
pretation, Kant’s view is a more sophisticated version of Hume’s. Like Hume, Kant would be rejecting the commonsense view that a mind is something distinct from the representations that it has.

There are at least two serious problems with this reading of Kant, even setting aside the metaphysical commitments of Kant’s moral philosophy. The first problem is that Kant certainly has the vocabulary to express a Humean bundle view, and yet instead typically expresses the relation of the self to representations in terms that would be hard for a Humean reading to accommodate. Perhaps most clearly: the self is the thing that combines representations in synthesis (cf. A77/B103). Kant nowhere suggests that representations themselves could have such a power.

The second reason why the Humean reading of Kant is implausible is that Kant is emphatic that we do not know ourselves as we are in ourselves. This is crucial, for instance, to his defense of the possibility of genuine free will (in the resolution of the Third Antinomy), and for his claim that we cannot know whether we are different in kind from the material objects we experience (in the Fourth Paralogism). What we are aware of are the appearances of ourselves. Appearances are representations, so if we are bundles of representations, then there needs to be another level of representations that is the appearance of that bundle. But this is clearly not Kant’s view—he holds that we can directly attach an ‘I think’ to any representation (cf. B131), and there’s no indication on his part that this amounts to the appearance of a representation.\footnote{45}

None of this rules out the (Spinozistic) possibility of the self’s (Brook 1994, 44). James Van Cleve thinks such a view is suggested by passages where Kant says that the self is never directly observed, on the basis of the further, Russelian claim that “logical constructions are always to be preferred to inferred entities” (Van Cleve 1999, 239) – an approach that could at most be taken as a suggested revision of Kant’s view.\footnote{45}

Kitcher grants that such worries show that her interpretation cannot reflect the whole of Kant’s views on the self, but insists that her interpretation is the most charitable one we can ascribe to Kant (e.g., “Kant vigorously rejects a phenomenal thinker, yet he cannot consistently advocate a noumenal one,” Kitcher 1984, 138, also Kitcher 1990, Chapter 5).

Kant’s Metaphysics of the Self

consisting of some accident or accidents of substances, so long as that accident or accidents are not the representations of which we are conscious.\footnote{46}

At the same time, Kitcher’s reading does have textual motivation. There are strong textual grounds (surveyed below) for believing that Kant thinks that there is an intimate connection between some interrelated set of representations and the self, and any plausible proposal concerning Kant’s positive metaphysics of the self must take this connection into account.

3.3 The self as a force?

Another interpretation of Kant’s views, inspired by his B423n. claim that the I “exists in the act,”\footnote{47} is to see the self as an act or activity. This relatively recent interpretation has been articulated by Rolf-Peter Horstmann and Arthur Melnick,\footnote{48} effectively tying Kant’s view to that of Fichte.\footnote{49} Both Horstmann and Melnick are drawn to this interpretation because it would give Kant a clear break with the metaphysical views of his predecessors.

While seeing the self as an activity appealed to some of the later German idealists, it is unlikely that Kant was drawn to it even in the second edition of the Critique, for the simple reason that he consistently ascribes activity to the self, without suggestion of identification. In the second edition of the Transcendental Deduction, Kant claims that “among all representations combination is the only one that is not given

\footnote{46} In the essay “What does it mean to orient oneself in thinking?”, Kant does claim that there’s something incomprehensible about the Spinozistic view of the psychological subject’s being an accident (“What does it mean” 8:143n.). Even there, however, he may be read as merely saying (along the lines of the Paralogisms chapter) that we cannot represent ourselves as an accident, where this stops short of implying any metaphysical conclusions.

\footnote{47} Guyer and Wood translate “in der Tat existiert” as “in fact exists.” The context does provide some reason for this translation—if they are right, then the present proposal loses its best piece of textual support.


\footnote{49} Fichte, for instance, says that “The I is, to begin with, nothing but a ‘doing’ (Fichte 1994, 81)
through objects but can be executed only by the subject itself” (B130), and such ascriptions are found throughout the critical works. There is simply no hint of an ontological shift towards seeing the self as a force or activity.\textsuperscript{50} Even the claim that the I “exists in the act” would be a bizarre way of identifying the self with an act: naturally read, it seems to say that the act is the condition for the self’s existence.

In addition, as we saw above, in the Paralogisms Kant says that the self might in some sense be one or more substances. Since there’s no indication that he thought acts could be substances, Kant couldn’t have consistently held that the self was an act or activity.

Of course, we would have some reason for suspecting Kant of beginning to shift towards an activity view if it were true that he could make his philosophy consistent only by construing the self as a force or act. But as we’ll see, that’s not the case. At the same time, Horstmann and Melnick are certainly right to see a fundamental connection between mental activity and our a priori understanding of the nature of the self,\textsuperscript{51} and any successful account will have to make sense of this connection.

3.4 The self as a form or structure?
Yet another position in the literature, inspired by claims such as “this I … is the mere form of consciousness” (A382), is that the I is a form. These interpretive positions are a little difficult to untangle, for sometimes they appear to be anti-metaphysical. For instance, Strawson claimed that, when it is separated from empirical bases of application, the representation ‘I’ merely expresses that unity which makes experience possible,\textsuperscript{52} and Günter Zöller claimed that most interpretations of Kant “show a remarkable agreement in their understanding of Kant’s thinking self as a form or structure that eludes any attempt at reification.”\textsuperscript{53} In the passage from Zöller, it is not clear whether this is some broadly Aristotelian view of the self as a form or structure, or whether the idea is closer to Strawson’s truly anti-metaphysical suggestion that the representation ‘I’ doesn’t refer to anything, but merely stands in a non-referring expressive relation to a form or structure.\textsuperscript{54}

But if we take this as a metaphysical view, how plausible is it as an interpretation? The first question to ask is whether the identification is with a type of form (e.g., a pattern of connection that various groups of causally independent representations might hold), or some token of a type. The former option seems very implausible, for it would imply that the self wasn’t a particular entity at all, but rather something that might be instantiated in various places. Kant never suggests anything so radical.\textsuperscript{55}

This leaves the possibility that the self is some particular instance of a form or structure. Yet this interpretive proposal would only face more serious versions of the problems faced by the previous candidate: Kant ascribes structure to us and our representations, and there is never any suggestion that a form could perform the sort of actions he ascribes to our mind (such as “synthesizing”). Moreover, Kant asserts that things in themselves must have non-relational natures,\textsuperscript{56} and

\textsuperscript{50} Kitcher considers a view along these lines and objects that the self “cannot be identified with acts of spontaneity, since these are distinct events” (Kitcher 1990, 122). This objection alone is not decisive, however, for the point of the view would be to identify the self with the single activity that is constituted by various instances of spontaneity.

\textsuperscript{51} Cf. Kant’s claim that “in every judgment I am always the determining subject of that relation that constitutes the judgment” (B407). Also see §25 of the B Deduction.

\textsuperscript{52} Strawson 1966, 167.


\textsuperscript{54} This would make ‘I’ more like “Goodbye” than like “this thinker.”

\textsuperscript{55} W.H. Walsh and others have claimed that something like this is what Kant means in talking about transcendental apperception (cf. Walsh 1997, 50). But even if this highly unnatural reading of the text is correct, it would leave the question at hand unsettled — for there certainly are individual selves, and the question is what Kant thought about their nature. That there is some universal that can be called a ‘self’ would be beside the point.

\textsuperscript{56} “[T]hrough mere relations no thing in itself is cognized” (B67). My claim is weaker than Rae Langton’s thesis that all properties of things in themselves are non-relational (Langton 1998).
forms are plausibly constituted by relations (cf. A20/B34), so his doctrine that the mind has an uncognizable nature in itself would become unintelligible. Finally, the best piece of textual support for this interpretation, the sentence at A382, is one that Kant revised in his own copy of the first edition, to say instead that the I is “the (unknown to us) object of consciousness”, and removed from the second edition.

4. An effect-relative compositional approach to the self

The previous section surveyed the obvious possibilities for what might be attributed to Kant as an a priori metaphysics of the self. The fact that there are clear problems with each of these candidates has certainly added motivation to the anti-metaphysical interpretive approach. If there is a plausible interpretation, it must be one that (1) allows some sense in which, for all we know, the self might be one or more substances or accidents, (2) keeps a clear sense in which we lack cognition of the self in itself, (3) allows the self to be outside of representations, and (4) maintains an intimate connection between the self, mental activity, and a certain unified set of representations that constitutes an experience.

I now argue for an interpretation that satisfies all four desiderata. Moreover, the view that I attribute to Kant is one that has independent philosophical plausibility, and so is one that can be charitably attributed to Kant. I first present the view in an intuitive way, and then turn to the textual basis for the attribution.

4.1 An intuitive approach

Consider what could plausibly be called a necessary truth about the mental representation ‘I’: any token of that representation will refer to whatever produced the representation, where ‘produced’ indicates a sort of immediate causal responsibility. There is no reason to doubt that Kant accepted this. We can then say that if some collection or

series of entities produced the token representation, then that group of entities is the referent. The core idea here is that we individuate the self (the referent of ‘I’) as the producer of a certain representation (the token ‘I’). If multiple entities were involved in the production of that representation, then we can say that they jointly compose the self.

We can extend this idea to cases where ‘I’ appears as part of some larger representation, e.g., the thought “I’m reading.” Say that in this case the “I’m” part of the thought was produced by a slightly different group of entities from that which produced the “reading” part of the thought (a physicalist might fill in the details here by saying that the thought was temporally drawn out enough that a few neurons died and were replaced while it occurred). It’s plausible to say that the self is constituted by whatever entity or entities produced the entire thought of which the token ‘I’ was a part, not just those responsible for the token ‘I’. That is, since ‘I’ is a singular term, we can say that the token ‘I’ refers not just to the various entities that produced it but to an individual composed of all the entities that produced the larger thought of which it is a part. Here, we individuate a self as the producer of an I-involving thought. To be sure, to avoid circularity, we have to assume that facts about what makes for a unified thought do not presuppose facts about antecedently unified or simple thinking entities, but that is not obviously absurd.

The next step is to eliminate the requirement that ‘I’ occur in the thought that individuates the self. Perhaps all that is required is that, with a little prompting, the entities producing the thought could have produced the thought with an ‘I’ in it. So a self could be individuated by a thought such as “This is getting interesting” just so long as, had things been slightly different, the entity or entities that produced that thought would have produced the thought “I think this is getting interesting.” And if that latter thought had been produced, then the ‘I’ it contained would have referred to an individual composed of all the

forms of entities.


58. The question here is at the heart of the argument from the unity of thought to the simplicity of the self, which was apparently endorsed by the pre-critical Kant. For a detailed discussion, see Ameriks 2000, Chapter 2.
entities that produced the larger thought. That an ‘I’ thought is possible in this way indicates that a certain amount of unified thought is already present.

The results of this approach depend on how big unified thoughts can get. If we can legitimately think of an entire course of experience as constituting a single thought (with smaller thoughts and perceptions as components), then we have a way of individuating selves that will line up closely with our everyday beliefs about how long and in what circumstances people ‘last’. Filling this out would require specific philosophical views about what constitutes a unified thought, but such views hardly seem impossible.

What sort of metaphysical view is this? One might be tempted to call it a response-dependent approach to the composition of selves. Yet we normally think that a response (in the relevant sense of ‘response’) presupposes a mind or self, so it would make little sense to use the normal notion of response-dependence here. Fortunately, there is a more general metaphysical principle that captures our intuitions.

The principle I want to appeal to states that a set of entities can compose a single individual on the basis of their jointly making a distinctive contribution to a single effect (in this case, a thought or experience). This metaphysical principle has independent intuitive appeal. For instance, we might speak of the production team for a film, where we know we’re referring to a non-arbitrary group of people simply because they jointly made a distinct causal contribution to some one thing. Other entities might have also contributed to the film (e.g., the actors), but because their causal contribution is distinct from that of the director, producer, and editor, they aren’t components of the production team. Likewise, even if other entities contribute to an experience (e.g., those that give rise to intuitions), that is a distinct contribution from that of the unifying activity, and so those entities aren’t part of the relevant individual (the self).

The operative metaphysical principle also surfaces elsewhere in the history of philosophy. An explicit statement of it comes in the seventh definition in Part II of Spinoza’s Ethics: “if a number of individuals so concur in one action that together they are all the cause of one effect, I consider them all, to that extent, as one singular thing.”

We can call this principle a principle of “effect-relative composition,” and the resulting view of the self the “effect-relative composition view of the self” (hereafter “the effect-relative view”).

I should address one potential confusion from the start. The effect-relative view is primarily a metaphysical view, not an epistemological one. So it does not follow from the fact that the self is composed relative to an effect that it is only known by inference from that effect. On the view I’ve described, it is true that we know ourselves only if we are producing some effect (some thinking), but we might say that we are equally immediately aware of the effect and of ourselves producing it (B157–58n. suggests this may have been Kant’s view). Metaphysical priority needn’t determine epistemological priority.

59. For a general response-dependent account of composition, see Kriegel 2008.

60. This approach has some similarity to Thomas Nagel’s suggestion that the self be understood to be whatever it is “in which the flow of consciousness and the beliefs, desires, intentions, and character traits that I have all take place” (Nagel 1986, 45). Nagel’s view is inspired by a Kripkean account of the reference of natural kind terms, however, and so doesn’t require that the metaphysics of the self itself necessarily involves facts about our mental life. Rather, our mental life is just what we use to refer to the self. For that reason, Nagel identifies the self with the human brain, whereas the view I’m presenting would at most say that the self is composed of a brain. Only the latter view allows that I could survive the death of my brain in cases where it has an appropriate psychological successor.

61. We might also say that all the entities that contributed to the effect in any way whatsoever constitute an individual. In the film case, we might call this the “film-making team.” Likewise, all the entities that contribute to a certain experience could be said to amount to an individual (perhaps an individual world—cf. §21 of the Inaugural Dissertation). But this is all quite compatible with there being some smaller individual that is the film-production team or the self.

62. There are some subtle questions here that would take some time to fully address, and it’s not part of my aim here to give a full account of Kant’s account of self-knowledge. With that said, I can tentatively offer one more level of epistemological detail: In being aware of oneself as thinking, on this view, one is immediately aware of some thing (or things) X that is producing a certain effect E. Now, X constitutes a self in virtue of producing E. So one is
Before turning to the textual basis for attributing a version of the effect-relative view to Kant, I should make it clear that the steps I’ve described here are not inferences. Someone could consistently accept the starting point of saying that ‘I’ refers to whatever produced that token mental representation and reject the rest of the picture I’ve described. But I do think they are natural steps, and that the resulting picture is very attractive, especially if we (like Kant) are certain that we’re something outside of our representations but aren’t certain what that something is.

4.2 Textual basis for attributing an effect-relative view to Kant

A careful look at the text shows that Kant in fact has both the specific concerns and the specific resources that would naturally lead him to an effect-relative account of the self, and that there are passages where Kant can plausibly be read as articulating a version of such an account. Together, these facts constitute a strong case for attributing some version of the effect-relative view to him.

Let’s start with the concerns. The concerns that motivate the effect-relative view focus on the possibility that, despite the simplicity of the representation ‘I’, what underlies thought cannot be assumed to be metaphysically simple (e.g., a Cartesian intellectual substance). Now, Kant undoubtedly holds that “messiness” on the level of the producer of thought can’t be ruled out a priori. The point of the Second Paralogism is that we cannot assume that a simple substance underlies each thought, and the point of the Third is that we cannot assume that an identical substance is present throughout changing states of thinking (though even there, Kant makes claims that clearly posit a single thing of some sort63). In the longer A version of the Paralogisms Kant illustrates this with physical analogies. As far as we know a priori, Kant immediately aware of the thing that composes or constitutes a self (albeit without knowing its intrinsic nature), and I take that to be sufficient for immediate awareness of a self. I’m grateful to an anonymous referee for pressing me to clarify this point.

63. E.g., “That the I of apperception, consequently in every thought, is a single thing that cannot be resolved into a plurality of subjects ... lies already in claims, the unity of thought might be the result of a collective unity of substances just as “the [single] movement of a body is the composite movement of its parts” (A353),64 and the continuity of representations might be based in a series of substances that pass their states from one to another, just as “[a]n elastic ball that strikes another one in a straight line communicates to the latter its whole motion” (A363n.). 65

Kant also has the resources needed for an effect-relative view. Such a view requires some theory of how mental representations could be unified into larger thoughts or representations, and Kant focuses much of his energy in the Critique on developing just such a theory. In fact, this is one of the central issues in the Deduction. Only a few paragraphs before introducing the notion of transcendental apperception, Kant states that a concept is a “single consciousness” that “unifies the manifold that has been successively intuited, and then also reproduced, into one representation” (A103), and he later explains how different perceptions are unified into a single experience:

There is only one experience, in which all perceptions are represented as in thoroughgoing and lawlike connection ... The thoroughgoing and synthetic unity of perceptions is precisely what constitutes the form of experience, and it is nothing other than the synthetic unity of appearances in accordance with concepts. (A110)

The discussion of how representations are unified is even clearer in the B version of the Deduction, whose opening section is entitled “On the concept of thinking ... but does not signify that the thinking I is a simple substance” (B407–08).

64. I see no good grounds for Ameriks’ claim that in the A edition Kant “preferred not to challenge the assertion of the simplicity ... of the soul as such” (Ameriks 2000, 51), though Ameriks is right that Kant’s attitude towards simplicity is less clear in his lectures on metaphysics.

65. Though he stops short of fleshing out the relevant metaphysics (perhaps due to his worries about Kant’s idealism), Robert Howell correctly notes that the elastic ball analogy shows that the identity of the self must be “formal” or “functional” (Howell 2001, 121, 141).
the possibility of combination in general” (B129), where Kant states that “among all representations combination is the only one that is not given through objects but can be executed only by the subject itself” (B130). Later, in (perhaps superficial) contrast to the A edition, Kant claims that judgments are the means by which representations are unified: “a judgment is nothing other than the way to bring given cognitions to the objective unity of apperception” (B141).

What we find in both editions, then, is Kant giving a theory of how representations are unified, which is exactly the sort of theory needed for an effect-relative approach to the self. I’ll call the version of the effect-relative view that uses Kant’s theory of how representations are unified the “Kantian effect-relative view”:

The Kantian Effect-Relative View. For any particular unified experience, whatever thing or things are immediately causally responsible for the unity of that experience compose a self.

The ‘immediately’ here makes room for the Spinozistic possibility that the self is an accident or accidents of some substance without being that substance, even though the causal activity of the accident may be ultimately attributable to the substance.

We’re now in a position to see that several passages can plausibly be read as statements of the Kantian effect-relative view. Here are three, with added italics to make the relevance clear:

[T]he original and necessary consciousness of the identity of oneself is at the same time a consciousness of an equally necessary unity of the synthesis of all appearances in accordance with concepts. (A108)

[T]he entire sensibility... must stand under universal functions of synthesis, namely of the synthesis in accordance with concepts as that in which alone apperception can demonstrate a priori its thoroughgoing and necessary identity. (A112)

Only because I can comprehend [the representations given in intuition’s] manifold in a consciousness do I call them all together my representations, for otherwise I would have as multicolored, diverse a self as I have representations. (B134)

So long as we’re not antecedently committed to the anti-metaphysical interpretation, each of these passages can be read as making the same, relatively straightforward point: in order for there to be a single self behind some set of representations, they must be unified (by an act of synthesis) into a single representation. In other words, Kant is telling us about the nature of the self, and is declaring that facts about the unity of representations are explanatorily antecedent to facts about the metaphysics of the self. Note that this view about antecedence is compatible with claims Kant makes in the Deduction which suggest that apperception is somehow the “ground” for the unity of representations (e.g., A117n., B134n.), for the latter claims can be read as saying that apperception has an epistemic, psychological, conceptual, or normative priority, not a metaphysical one. For these same reasons, the

66. If we stay true to the text, this will be noumenal, atemporal causation. Atemporal causation has struck many as incoherent, but the idea of it is hardly unique to Kant—cf. Augustine’s view of God’s relation to the temporal universe.

67. Contrary to Kitcher 2006, 199, there is no incompatibility between the idea that the self constructs knowledge and the idea that it constructs itself. Rather, it is by constructing knowledge that self constructs itself. And contrary to Kitcher 1990, 122, seeing the self as what’s responsible for synthesis doesn’t require seeing synthesis as a conscious, intentional activity.

68. On a related point: one might worry that the effect-relative view sits uncomfortably with Kant’s claim that identity of subject across representations “does not yet come about by my accompanying each representation with consciousness, but rather by my adding one representation to the other and being conscious of their synthesis” (B133). While we might read this as saying that an identical subject is metaphysically prior to the identity of self-consciousness in representations, it could also be taken as saying that there must be a unified act of synthesis prior to there being identity of consciousness. So long as a unified act of synthesis does not presuppose an antecedently unified
Kantian effect-relative view is also compatible with Kant’s claims that we cannot think of ourselves in the same way we think of objects (e.g., A346/B404, B422). Because of this, the view is consistent with, and has the potential to clarify, a range of interpretations of the Analytic.

In the above passages, Kant specifies which of the entities that contribute to a thought count as constituents of the self. It is those entities that contribute to synthesis, the combining of representations. Other entities give rise to sensations (uncombined, “raw” representations), and those entities don’t count as constituents of the self. In the hylo-morphic terms Kant sometimes uses, it is those entities that generate the form of an experience that compose a self. In more contemporary terms, we might say that the self is composed of whatever entities are responsible for processing data into an experience, not those that generate the data.69

This interpretation satisfies all the desiderata listed at the beginning of this section. Once again, we want a view that (1) allows some sense in which, for all we know, the self might be one or more substances or accidents, (2) keeps a clear sense in which we lack cognition of the self in itself, (3) allows the self to be outside of representations, and (4) maintains an intimate connection between the self, mental activity, and a certain unified set of representations that constitutes an experience.

On the effect-relative view, we do not know what is immediately causally responsible for unifying representations into a single experience. It could be several substances or accidents, and we have no

69. The effect-relative view actually provides a way of fleshing out many of the merely suggestive claims about Kant’s theory of the self that are found in the literature. Apart from his view of the self as an act, this can be said of Horstmann’s proposal that the self be understood on analogy with a center of gravity (in other words, a functional-role account of the self) in Horstmann 2007, and of Richard Aquila’s proposal that “the purely ‘formal I’ is a ‘determining’ form or structure whose original embodiment in, or determining of, the I of inner sense is what constitutes the latter as a ‘soul,’ and thus as a thinking being or an ‘I’” (Aquila 1997, 34).

insight into their nature apart from their contribution to that experience. This means that the view satisfies desiderata (1) and (2). Nevertheless, the effect-relative view does take the self to be composed of some entity or entities that are outside of its representations, and so it satisfies (3). Last, the view asserts that the relevant entity or entities compose a self in virtue of their unifying representations into an experience. So it also satisfies desideratum (4). 70

5. Philosophical virtues of the Kantian effect-relative view

The Kantian effect-relative view is a relative of Locke’s view of personal identity.71 Locke’s view has, by and large, had a positive reception. Much of the appeal of the Lockean view stems from its metaphysical modesty: it requires no assumption that the self is fundamentally different from material beings, and can allow for personal identity to get “messy” in imagined cases of people fusing and fissioning. The Kantian effect-relative view should have a similar appeal, though much will hang on how acceptable the Kantian understanding of the structure of representations is—that is, on whether we can find an understanding of ‘synthesis,’ ‘intuition,’ etc., that fits into the ontology of contemporary views.72

But the effect-relative view also has virtues that the (original) Lockean view lacks. It imposes a stricter requirement on two entities’ being part of the same self—a mere memory link isn’t enough, but rather

70. There is a further complication that the view should perhaps be extended to address. Kant holds that, for all we know, inner and outer sense may have the same ground (cf. A379–80, B409). This in turn opens up the possibility that different selves might have the same ground (cf. Melnick 2009, 38). To maintain the distinction between the different selves, then, we may need a principle of effect-relative division, according to which one thing can constitute several things on the basis of producing several effects (in this case, experiences). Elsewhere, I hope to explore this further principle, and its relevance to Kant’s broader philosophy.


72. A number of interpreters of Kant have worked to do precisely this. Kitcher’s work is one example. See also Meerbote 1990. Allison 1996a resists just such attempts, but is based in an anti-metaphysical reading of Kant. Kant himself clearly holds that at least our faculty of reason cannot be naturalistically explained (cf. Groundwork 4:452, Critique of Judgment 5:254).
a sort of “experiential continuity.” It hasn’t been part of this paper to delve into Kant’s views on what makes for a unified experience, but it clearly would involve much more psychological continuity than single memory links. This is a virtue of the view—having a single memory passed from one being to another seems like too little for them to be parts of the same self.\textsuperscript{73}

The Kantian effect-relative view also ties certain types of mental activity to the survival of the self, specifically, those activities involved in maintaining a unified experience. This provides a striking egoistic justification of our general concern with having a coherent picture of the world. On the view I’ve described, we express something close to the literal truth when confusion and scatter-mindedness leads us to say that we’re “falling apart.”

Last, the Kantian effect-relative view, taken as something a priori, offers an explanation of how the representation ‘I’ is used without the worry of reference failure, while still seeming, as Kant says, “empty”. ‘I’ involves no determinate conception of the self, and so could be wielded in scenarios such as Anscombe’s sensory deprivation tank.\textsuperscript{74} It gives an explanation of how, in such a scenario, we might know that we’re the same person who stepped into the tank even though we’ve lost all sense of our bodies. For as long as we are aware of a unified course of experience that stretches from the earlier time to the present time, we have a grasp on what we need to individuate our self.

Of course, much more detail needs to be provided in order for the view to be a contemporary contender for how we understand the self. I suspect that some of the details of Kant’s actual view won’t be plausible to many contemporary philosophers in the end—if nothing else, because Kant’s theory of representations posits a deep contrast between us and non-rational creatures.\textsuperscript{75} But some variant of his view may be a serious contender, for it allows us to have a priori certainty that we exist and some a priori knowledge of our nature while being agnostic concerning our fundamental constituents.

**Conclusion**

The account I’ve described raises a number of interpretive and philosophical questions, including: What exactly is involved in the crucial unified experience on which the Kantian effect-relative view is based? How does the view allow for Kant’s realism about and assumed insight into mental faculties? How far back in the causal chain can the constituents of the self be, and how does it make the cut-off (so that one’s parents are not part of one’s self)? How does this view of the reference of ‘I’ tie in with our conception of ourselves as embodied entities inhabiting the physical world? And especially: What methodology did Kant think entitled him to this view of the self?

These are important questions. Some speak to issues that are specific to my interpretation, while others stem from general interpretive or philosophical issues. I’m confident that some have relatively clear answers, but I’m unsure about the others.\textsuperscript{76} Nevertheless, the Kantian effect-relative view allows us to clearly address issues in Kant’s work that have been obscured by the dominance of the anti-metaphysical interpretation.\textsuperscript{77}

\textsuperscript{73} Cf. Parfit 1984, 206. Kitcher notes a similar virtue for the view she attributes to Kant (Kitcher 1990, 124–25). There is, of course, a large literature on Locke’s view and personal identity that I’m not touching on.

\textsuperscript{74} Anscombe 1975. Howell 2001 discusses how particular references to the self might work, but doesn’t address issues of keeping track of oneself across times. Note that it is because the effect-relative view is supposed to be \textit{a priori} that it can help explain self-reference and the self-ascription of representations. This is in contrast to, say, Kitcher’s Humean view of the self, which is presented as an inference to the best explanation (Kitcher 1990, 125).

\textsuperscript{75} Cf. Metaphysik Mrongovius 29:906, Metaphysik Vigilantius 29:1027, and Anthropology 7:127.

\textsuperscript{76} On the question of methodology: there are at least three prima facie plausible accounts of how Kant would have arrived at this view: it could be result of conceptual analysis, it could be a given “fact of reason” (cf. Critique of Practical Reason 5:31), or if having a self is a necessary condition for experience, it could even be the result of a transcendental argument.

\textsuperscript{77} For much helpful feedback, I am grateful to Béatrice Longuenesse, Don Garrett, David Velleman, Karl Schafer, John Richardson, Jeff Sebo, John Morrison, Jonathan Simon, Johannes Haag, Mike Raven, Dina Edmundts; the participants of New York University’s 2007–2008 Thesis Preparation Seminar; Rolf-Peter Horstmann and the members of his 2009 colloquium at Humboldt Universität; and two referees for Philosophers’ Imprint.
Works Cited


Kant, Immanuel. 1900-. *Kants Gesammelte Schriften*, German Academy of Sciences (ed.). Berlin: De Gruyter.


Sellars, Wilfred. 1970. “…this I or he or it (the thing) which thinks…”. *Proceedings of the American Philosophical Association* 44, 5–31.


