Aporia vol. 35 no. 1–2025

Designation and Representation: Understanding Kant's Transcendental Unity of Apperception

Elizabeth C. M. Hornsby

Kant launches an attack on what he calls "rational psychology" in the Transcendental Dialectic in the chapter entitled "The Paralogisms of Pure Reason." The Paralogisms identify claims made in traditional "dogmatic" metaphysics about the soul as properly lying outside of the bounds of pure reason and seek to explain how the fallacious identification of certain features of the soul have come about. In doing so, they bring to light the unique nature of Kant's doctrine of the "transcendental unity of apperception" in contrast to traditional conceptions of our acquaintance with the self or soul. In the following discussion, I will focus in particular on the aims of the first three Paralogisms, as the fourth is better understood as relating to Kant's own Transcendental Idealism than rationalist psychology. The topics of the first three Paralogisms are the soul as *substance*, as *simple*, and as *persisting* (i.e., identical with itself over time).

Through his analysis of these issues, Kant brings to light a common theme in his predecessors' philosophy: the false assumption that we have a true "representation" of the soul from which we can make inferences as we would for any other representation. What we actually have is some other species of mental item altogether. I will begin my argument by establishing what and whose position the Paralogisms intend to defeat. I will then argue for an interpretation of the error Kant attributes to the rationalists based on a distinction between *designation* and *representation* which is only briefly alluded to in the Paralogisms, but which is key to understanding

Elizabeth C. M. Hornsby is currently reading philosophy and German at Oxford University and will graduate in July 2025. She intends to continue her study of philosophy at the graduate level, focusing broadly on ethics, epistemology, and Kantian philosophy.

both their content and the transcendental unity of apperception. Finally, I will evaluate the success of this newfound understanding of the soul on its own merit: regardless of whether this interpretation accurately represents Kant's own view, is it one we ought to accept?

1. The Rational Psychologists

The targets of Kant's attack on rational psychology are rationalists, whom Kant describes as "dogmatic" metaphysicians (B23). Chief among these are Leibniz and Descartes; it was within the Leibnizian–Wolffian tradition that Kant himself was educated, and Descartes' *cogito* would have been a prominent moment in the rational philosophy of the soul which could not have been ignored (Schönfeld and Thompson). The rationalist school of metaphysics, insofar as it can be considered as a unified endeavour, is often defined in contrast to the empiricist school: whereas empiricists believe that all knowledge proceeds from experience, rationalists attempt to derive knowledge from what Kant would call "pure" sources, i.e., completely *a priori* principles. Thus "rational psychology" is the attempt to attain substantial knowledge of the nature of the soul by means of *a priori* reasoning alone. That this is the kind of dogmatic metaphysics Kant attempts to disprove in his Paralogisms is evident from various points in the text, perhaps nowhere more explicitly than when he writes:

> **I think** is thus the sole text of rational psychology, from which it is to develop its entire wisdom. One easily sees that this thought, if it is to be related to an object (myself), can contain nothing other than its transcendental predicates; because the least empirical predicate would corrupt the rational purity and independence of the science from all experience. (A343/B402)

On the account of a rationalist such as Descartes, "I think" is indeed the sole text from which all else must proceed, the one indubitable item of knowledge from which he claims to be able to deduce the rest of his doctrine about the soul (and indeed about everything else). This thought contains nothing other than its "transcendental" predicates, where "transcendental" means its property of being a necessary condition of possible experience, because any empirical predicate would corrupt its "rational purity and independence of the science from all experience" (A343/B402). Any empirical notion of the thinking subject would not be appropriate for the kind of theory he is aiming at where the thinking subject must be thought of in *pure a priori* terms only.

The worry might then be raised whether any of Kant's Paralogisms, which are supposed to be attributed to his rational psychologist opponents to show the error in that line of thinking, actually represent the arguments made by the likes of Descartes, Leibniz, and Wolff. Patricia Kitcher argues that the Paralogisms ought to be regarded as "very much Kant's own Paralogism[s]" and that they are better suited to representing caveats to his own arguments than to defeating arguments made by the rationalists (531). In particular, she argues that the minor premises of the first three Paralogisms are only acceptable to Kant based on an understanding of his own doctrine as laid out in the Transcendental Deduction. This would suggest that the inferences of the Paralogisms cannot properly be attributed to the rationalists because part of their acceptance presupposes Kant's own critical philosophy.

However, this view rests on a flawed understanding of what a Paralogism is. Ian Proops' textual analysis of Kant's various definitions and uses of the term concludes that the following criteria are necessary for a syllogism to be considered a Paralogism: (1) that it is mistaken with regard to form: specifically, that it has an ambiguous middle term; (2) that the premises are correct; and (3) that the error is tempting to make—in this case, because it rests on a transcendental illusion (470). If the middle term—whatever it may be—is ambiguous, this suggests that for both the major and the minor premise, there could be multiple possible readings and hence multiple possible versions that could be accepted on different grounds. That Kant's own acceptance of the minor premise is based on his novel insights from the Transcendental Deduction does not imply that no alternate construal of that same premise could have been accepted by the rationalists on entirely different grounds. For example, the minor premise of the first Paralogism (as in the B edition) is:

> Now a thinking being, considered merely as such, cannot be thought otherwise than as subject. (B411)

Certainly, some version of this premise can be attributed to the likes of Descartes, whose notion of substance—the bearer of properties, abstracted from those properties which inhere in it—is closely related to that of subject, which is the grammatical equivalent, the thing predicated in a judgment. Descartes' slide from the premise "I think" to "I am a substance whose essential property is to think" can be thought of as following something like Kant's first Paralogism.

Similarly, Kitcher cites an interpretation of Leibniz's thinking machine thought experiment as a rationalist model for Kant's second Paralogism, that of simplicity. Margaret Wilson explains that this thought experiment shows Leibniz's view that the unity of thought could never

come about through a complex machine, and that a thinking subject must thus be simple (Kitcher). That this particular passage in Leibniz might have inspired the second Paralogism could serve to explain the unusual definition of "simple" Kant gives in the major premise of this argument ("that thing whose action can never be regarded as the concurrence of many acting things, is simple" (A351)), but, more importantly, it lends plausibility to the idea that some rationalist might have accepted some construal of the minor premise of the second Paralogism (as in the A edition):

> Now the soul, or the thinking I, is such a thing [whose action can never be regarded as the concurrence of many acting things]. (A351)

Though the Paralogisms do not occur as Kant wrote them explicitly in the writings of the rational psychologists, they are best understood, as Longuenesse argues, as displaying the "implicit logical structure" behind the arguments to which they appeal (21). Since it is indeed plausible that these are assumptions made by the likes of Descartes and Leibniz, the objection that Kant's attack on rational psychology misrepresents his opponents does not hold water.

This interpretation implies a minor but noteworthy consideration about the kind of error that Kant would be attributing to the rational psychologists. We know that the error in the Paralogisms is a formal one consisting in the ambiguity of the middle term, and thus that the versions of the major and minor premises that Kant accepts respectively do not together form a valid argument leading to the conclusion. However, if one assumes that the rationalists accept a different version of the minor premise than Kant—because they have not achieved critical philosophy yet—then the error they make in accepting the Paralogisms cannot itself be formal because they endorse versions of the major and minor premises such that the structure of the argument is valid. Instead, the error would be in accepting that version of the minor premise in the first place—an error which comes about due to the ambiguity of the middle term and the transcendental illusion of taking the subject of the transcendental unity of apperception, the "I" in "I think," to be a given object, which illegitimately enables the move from Kant's version of the principle to the version attributed to the rationalists.

The Paralogisms are paralogisms in the sense that one who endorses true versions of the major and minor premises must recognize that the argument is not valid due to the double meaning of its middle term, but not in the sense that the versions of the arguments the rationalists actually endorsed were invalid (they were merely unsound). The rationalists have fallen victim to the

transcendental illusion *before* making their arguments, as it is the cause of their accepting the wrong versions of the premises.

2. "I think" as Designation

But what is the difference between Kant's version of the premises and the rationalists'? How does Kant interpret the minor premise, and how do the various explanations of the error—transcendental illusion, wrongly treating the representation "I" as if it contains a manifold and an ambiguous middle term—come together? To answer these questions, it will first be helpful to examine the Paralogisms as they are presented in the text.

In the A edition, Kant gives four Paralogisms in full syllogistic form, of which the first three are most relevant to this essay; in the B edition, only one is written out. Its content corresponds to the first A edition Paralogism but as the only one remaining from the first edition of the *Critique*, its form must be understood as exemplifying the mistake made in all of the Paralogisms before it:

> What cannot be thought otherwise than as subject does not exist otherwise than as subject, and is therefore substance.
>
> Now a thinking being, considered merely as such, cannot be thought otherwise than as subject.
>
> Therefore it also exists only as such a thing, i.e., as substance. (B410–411)

This argument is supposed to reflect the rationalists' method of psychology which enables them to move from the representation of the self as subject to the real properties of the self as substance. Kant is not accusing his predecessors of a facile or simple error here; the argument does seem to be valid, so it is a genuine question Kant is answering in his criticism of the Paralogisms: wherein lies the problem? According to Proops' analysis of the term Paralogism, the error is formal, and therefore Kant's qualms with the argument must not consist merely in his dismissal of either of the premises (470). However, Buroker construes him as outright rejecting the major premise (216). Her argument stems from the distinction Kant makes between dogmatic, skeptical, and critical objections at A389. Buroker argues that Kant rejects dogmatic knowledge of the major premise as "the critical position claims that 'the assertion is groundless, not that it is incorrect'" (216). However, this misidentifies which assertion the critical

position is said to be aiming at. Since the critical position is aimed at the "proof" of a proposition rather than the proposition itself, it would be more consistent with Kant's classification system to identify the problem with the Paralogism as being found within the structure of the inference rather than the acceptability of either of the premises. His critical objections leave the proposition "untouched in its worth or worthlessness": it is important to his methodological commitments in the Paralogisms that he does not need to deny either premise to deny that we know the conclusion (A388). The discussion under the heading "Criticism of the first paralogism of pure psychology" confirms this in that it shows what Kant intends to do instead. Here, he does not explicitly deny the major premise but rather emphasizes that the "I" of "I think" is a logical subject rather than a given object, and hence suggests not that the major premise is false, but that it does not properly apply to this representation of "I." This is an entirely different type of problem with the Paralogism than the one Buroker reads into it: it is a problem not with the argument itself, but with the way that the first premise must be stretched to breaking point to accommodate the "I" of the transcendental unity of apperception. Seeing as the evidence against it is insufficient, I will follow Proops in taking the major premises of the Paralogisms to be definitional truths with which Kant agrees and consider further the sense in which he endorses the minor premises.

Extracting the key insight of the Paralogisms into the transcendental unity of apperception is a task bound up with identifying what exactly the ambiguous middle term is, as well as what its two different readings are. Given that Kant himself equivocates on what that term is—at B411 it is the "being" which is mentioned in the minor premise and implied in the major; at B411–412 in a footnote, it is the "thinking"—this is no easy task. As both passages pick out words used imprecisely, it is more fruitful to look for an ambiguous concept than an ambiguous term which serves as the middle. In both the main text and the footnote, the key distinction being drawn is between something "as it might be given in intuition" and something which is only thought in relation to self-consciousness. Something that might be given in intuition is an object—something which has content that can be cognized, hence why its contrast in this context is referred to as the "*form of thinking*" (emphasis mine).

The connecting thread that underlies Kant's discussion of all of the Paralogisms in both editions of the text is the fallacy of treating the representation of the self in the "I think" as if it were something that could be cognized, or as if it contained some given content—as if it is a representation, essentially. He states numerous times quite categorically that it is not, in fact, a representation, in the sense that he has defined it throughout the *Critique*. In the criticism of the first Paralogism, he writes, "apart from this

logical significance of the I, we have no acquaintance with the subject in itself that grounds this I as a substratum" (A350); in the second, he calls this expression "wholly empty of content" (A355); in the B edition, he emphasizes that no object is cognized merely by the fact of thinking (B407). This expresses a point which was already present in the Transcendental Deduction—that the "I think" which must be able to accompany all my cognitions does not, itself, contain a manifold; it is not an intuition (B138). The temptation would be to classify it as a concept if it is not an intuition, especially since it is related to thinking, the function of the understanding, which deals in concepts—but Kant makes explicit at A382 that it is not a concept either. He explains that it is "the mere form of consciousness," but this does not entirely clear up what kind of representation this "I" is.

I propose that the fallacy consists in the fact that the "I" is not, properly speaking, a representation at all. It cannot be one, since it does not bear the relation of representation to any object. An *intuition* of a book "represents" a book because it is a matter of the subject being given a book as object through the senses; the *concept* of a book "represents" the object in a slightly different way in that it contains within it more specific concepts which make up the characteristics of a book, such that through this more general concept, the empirical intuition of a book can be identified as such. Call the relation of representation to object *R*, where R_i is the relation as it is in intuition and R_c is the relation as it is in concepts. What R_i and R_c have in common is that the object they are representing features in the causal chain of the representation coming about; the difference is merely that one is a mediate and one an immediate representation, and so the *type* of *R*-relation specifies whether there are more or fewer steps involved in that causal chain between the object and its representation. Could it be possible that the relation that the "I" of the transcendental unity of apperception bears to the thinking subject in itself is another species of *R*-relation (say, R_{tua})? No: given what I have identified as the core characteristic of the R-relation genus, the causal connection between the object being represented and the representation of it, there can be no R_{tua} as there is no causal link between the subject in itself and the "I" of the transcendental unity of apperception.

Neither the intuition of an *x* nor the concept of an *x* can arise, ultimately, without the thinking subject having an experience of some *x*, but while the transcendental unity of apperception cannot arise without *some* experience—this is what Kant means when he says it is *abstracted* from all experience as opposed to *separated* from—it is object-indeterminate; the "I" of the transcendental unity of apperception does not require experience of the thinking subject, as the subject in itself is never and can never be given (B427). That this is the crux of his insight is evident at A381 where he writes

that this representation is simple "only because [it] has no content, and hence no manifold, on account of which it seems to represent a simple object, or better put, it seems to *designate* one" (emphasis mine). Classifying the "I" of the transcendental unity of apperception as a *designation* rather than a true *representation* of an object emphasizes its lack of *R*-relation to the object it designates. Hence, Kant is able to justify the observation that what principles might hold in general for inferences drawn from representations to their objects—such as the major premise of the first Paralogism, for example—do not need to hold for the equivalent inference from designation to object. After all, the categories, including the notions of substance, simplicity, and identity which are in question here, only have legitimate use for the objects of possible experience, a set to which the subject in itself does not belong.

This distinction gives clarity to a conclusion of Kant's which was otherwise quite obscure. He writes: "One can quite well allow the proposition **The soul is substance** to be valid, . . . [but] it signifies a substance only in the idea but not in reality" (A350–351). It is difficult to make sense of this statement when we use the language of representation. If I represent some *x* as having a certain quality *y*, then in saying "*x* is *y*," I can correctly be understood to be attributing the quality *y* to both my representation of *x* and the given object *x*—this is the purpose of a representation and a natural consequence of the fact that a representation of *x* is related in the appropriate way to the object *x*. But for the transcendental subject, the "I" of the transcendental unity of apperception, these two things seem to come apart. Kant wants us to be able to say that "The soul is substance" is true if we are attributing the quality of substantiality to the representation we have of the soul, but not if we are attributing it to the soul itself as object. The disparity between his treatment of representations in general and the "representation" of the soul is another clue that the latter is not really a representation at all. After all, as he writes, "this concept of ours leads no further . . . it cannot teach us any of the usual conclusions of the rationalistic doctrine of the soul" (A350–351). It can lead us no further because it *goes* no further: dig deeper into a representation and you will find the object it is a representation of; dig deeper into the designation of the transcendental subject and you will find no such thing.

Neither Buroker nor Kitcher is thus incorrect in interpreting the minor premises of the Paralogisms as consequences of a uniquely Kantian doctrine of the transcendental unity of apperception. This interpretation—where the "I" is a designation rather than a representation—is correctly identified as the one which Kant endorses, but it is not the one he ascribes to his rational psychologist opponent. The mistake made by the rationalist is supposed to be the misidentification of "I" as a true *representation* and treating the necessary properties of the *designation*—that it is really

substantial, simple, and identical with itself—as if they were properties of an object to which it bears the relation of representation, when it in fact does not bear this relation to any object. Kant can therefore agree that the major premise is true of any representation of an object but that the minor premise does not concern only a designation of an object, and so the argument is not valid. The discussion from the B edition footnote which cites the term "thinking" as the ambiguous middle term of the syllogism, though obscure, can be read as supporting this argument:

> "Thinking" is taken in an entirely different signification in the two premises: in the major premise, as it applies to an object in general (hence as it may be given in intuition); but in the minor premise only as it subsists in relation to self-consciousness, where, therefore, no object is thought, but only the relation to oneself as subject (as the form of thinking) is represented. (B411)

This interpretation helps explain the distinction previously identified: it is a distinction between a major premise which applies to *objects* or *things*, and a minor premise which does not—it only talks about thinking as a relation to oneself, as abstracted from every object, or as purely form rather than content. It is a mistake to identify this transcendental construction as an object.

The final piece of the puzzle is the concept of transcendental illusion. Transcendental illusion is, according to Kant, an inevitable illusion caused by the conflation of subjective with objective grounds of judgment. Reason follows the legitimate principle expressing its task, "to find the unconditioned for conditioned cognitions of the understanding, with which its unity will be completed" (Buroker 209–210). This principle is subjectively necessary because it is regulative for human reason, and the transcendental illusion comes from confusing it with an objective principle: "when the conditioned is given, then so is the whole series of conditions" (209–210). In the case of rational psychology, the influence of the transcendental illusion can be seen as the motivating force behind the move from Kant's accepted version of the minor premises (which are merely consequences of the transcendental unity of apperception) to the versions he ascribes the rational psychologists (which wrongly represent the "I" of transcendental apperception as a thing or object given to us).

The "unconditioned" that the reason is tasked with finding would in this case be the absolute subject of all cognitions—the subject in itself, as designated by the "I"—and the transcendental illusion consists in the mistaken assumption that, given the reason is charged with searching for this unconditioned, it must be given just as the conditioned cognitions are given. This leads to the misunderstanding of the purely formal features of

the "I" designation as being objective features of the self that it designates. The nature of transcendental illusion is that it cannot merely be explained away. Like optical illusions, it is persistent, remaining even once one rationally recognizes that it is an illusion and requiring continuous effort of reason to dispel (A296–297/B353–354). Kant is not, therefore, accusing his predecessors of a simple or obvious mistake, but grants that making such a mistake is inevitable given a faculty of reason which has not yet been subject to critique. Rather, he admits that the properties of substantiality, simplicity, and persistence do necessarily belong to the designation "I" of the transcendental unity of apperception; it is a natural yet nonetheless fallacious assumption that the same properties apply, in a different sense, to the object which it designates.

3. Conclusions

With all these pieces in place, we can evaluate whether this position is viable or not. Given (1) that his argument does identify a real difference in meaning between two versions of the premises in the Paralogisms, leading to their being subtly but fatally formally invalid on what Kant would consider the "correct" reading of each premise and (2) that the ideas behind the Paralogisms are correctly attributed to the rational psychologists, even if they are not to be found word-for-word in rationalist doctrines, there is only one remaining way that the argument could be shown to be unsuccessful. If Kant's idea of the "correct" reading of the minor premises is wrong and the rational psychologists are in fact entitled to the version of the minor premise which posits the subject as a representation of an object rather than a formal construction or designation, then this would complete the valid version of the syllogism and entitle them to its conclusion as well. Kant has shown that his designation of "I" does not entail the representation version of "I," so if the rationalists really do proceed from the same point as him and take the step from merely "I think" to there being a given thinking being, then they are subject to the transcendental illusion. But if they were to acquire a representation of the self by other means, for example, as a given object (an intuition), then they could be entitled to the version of the premise which would make the argument valid.

Thus the question becomes whether or not there is or could be an intuition of the self. Kant says that there is none, and he seems to be correct. Certainly for Descartes the move seems to be straight from "I think" to "I am a thinking substance"; the direct acquaintance is with the perceptions and thoughts which he considers properties of the substance, not with the substance itself. But it is not just Descartes who thinks this.

Any committed rationalist would have a hard time disagreeing, as to found their psychology on an intuition of the self would be to abandon the idea that these facts about the soul can be known purely *a priori*, without reference to empirical experience. Despite Kant's parting company with Hume with regard to the unity of mental items and the existence of a thinking subject, it seems he ultimately follows him in the insight that led to his bundle theory: there is no occasion where an unchanging substantial "self" is given to us as an object, only individual perceptions.

Kant's aim in this section of the *Critique of Pure Reason* is about more than simply demonstrating that the contemporary rationalist metaphysics of the soul were flawed; it is about coming to a substantially new and substantially *better* understanding of the soul as something we cannot *represent* it in the same way we can represent any other object. It is the role of the faculty of reason to provide explanations—to move from the conditioned to its conditions (A299–300/B356–357). Thus it is the assumption that wherever there is conditioned, there must be, and we must always be able to discover, conditions. This motivates the Paralogisms and provides the force behind the move from "I" as designation (or logical construction) to "I" as representation, intuition, or object. Since, according to Kant, it is an inevitable fact of human nature that we are fooled by transcendental illusion. Even when we know rationally that it is an illusion, the presence of this illusion in the Paralogisms is relevant not only to the rationalists but to everyone, as it is the kind of error human reason is naturally prone to and must be critiqued to keep in check.

This provides both a reason for the *Critique* and a warning for the critical philosopher not to become complacent. Understanding this chapter not only in terms of its attack on rational psychology but more generally as an antidote to transcendental illusion helps clarify the observation made earlier about the form of error attributed to the rationalists. They had already fallen victim to transcendental illusion prior to making the arguments represented in the Paralogisms, and their reading of the Paralogisms was *unsound*, but not *invalid*. A doubt that remained about this interpretation was whether it really fit with the definition of paralogism as a *formal* failure in a syllogism if the error being ascribed to the rationalists did not consist in form at all. However, the formal error is still there: only instead, it is an error that the *critical philosopher themself* is tempted to make by the irresistible nature of the transcendental illusion. The Paralogisms are simultaneously a corrective for the rationalists *and* for the critical philosopher.

Works Cited

Buroker, Jill Vance. *Kant's "Critique of Pure Reason": An Introduction*. Cambridge UP, 2006.

Kant, Immanuel. *Critique of Pure Reason*. Translated by Allen W. Wood and Paul Guyer, Cambridge UP, 1998.

Kitcher, Patricia. "Kant's Paralogisms." *The Philosophical Review*, vol. 91, no. 4, 1982, pp. 515–547.

Longuenesse, Béatrice. "Chapter 1. Kant's 'I Think' versus Descartes' 'I Am a Thing That Thinks.'" *Kant and the Early Moderns*, Princeton UP, 2008, pp. 9–31.

Proops, Ian. "Kant's First Paralogism." *The Philosophical Review*, vol. 119, no. 4, Oct. 2010, pp. 449–495.

Schönfeld, Martin, and Michael Thompson. "Kant's Philosophical Development." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2024, Metaphysics Research Lab, Stanford University.