To know the self as a matrix of maybe:
An account of the specialness of self-knowledge
# Table of contents

Chapter 1. Introduction ......................................................................................... 3  
  1.1 A brief overview of this essay .................................................................. 3  
  1.2 The specialness of self-knowledge ......................................................... 3  
  1.3 Method ....................................................................................................... 7  

Chapter 2. The Self.................................................................................................. 11  
  2.1 Moore’s distinction and no-self views .................................................. 12  
  2.2 The self and the singular first-person pronoun .................................... 16  
  2.3 The Writer from the observer perspective ............................................. 20  
  2.4 The Writer from the participant perspective ......................................... 27  
  2.5 Bridging the perspectival gap ................................................................. 42  
  2.6 The self as a reflective matrix of maybe .............................................. 47  

Chapter 3. Self-knowledge .................................................................................... 55  
  3.1 Knowledge .................................................................................................. 55  
  3.2 What it means to know yourself .............................................................. 61  

Conclusion............................................................................................................. 69  

References ............................................................................................................ 71
Chapter 1. Introduction

1.1 A brief overview of this essay

The goal of this essay is to make sense of the apparently special relation between self-knowledge and agency. To achieve that goal, the essay translates the account offered by Sartre (1994) of what it is like to be a human self into the language of evolutionary psychology. Sartre describes the phenomenology of the self as a series of inescapable choices in a contingent set of circumstances. This essay identifies Sartre’s description with what Baumeister et al. (2018) call a matrix of maybe: the mechanism of nonfactual pragmatic prospection found in humans. Consequently, it defines the self as a matrix of maybe operating within a contingency matrix and reflecting on its own operation. Self-knowledge, the essay concludes, seems special because we routinely and erroneously ascribe to the self features of its contingency matrix. Most of our true first-person claims should not be read as I PREDICATE. Instead, they can be explicated as I have to act in a world where C PREDICATE, where C is the relevant part of the contingency matrix.

The essay consists of three chapters and a conclusion. In chapter 1 (“Introduction”), I clarify what I mean by the specialness of self-knowledge and discuss the method of my investigation. In chapter 2, which forms the bulk of the essay, I define the self in six consecutive steps; there is a summary of each step at the beginning of the chapter. In chapter 3, I adopt a definition of knowledge and then consider the implications of my view of the self in light of that definition. The conclusion sums up the results of my investigation in jargon-free language and suggests ideas for future exploration.

1.2 The specialness of self-knowledge

My goal is to make sense of self-knowledge. It is reasonable to ask what there is to make sense of. In what way does self-knowledge pose a special problem? What, in other words, makes it different from other knowledge?

In recent Anglophone philosophy, there have been two main approaches to self-knowledge (see e.g. Gertler 2010 for an overview). Roughly, one approach focuses on our epistemic
access to our mental states. The other foregrounds the relation between our mental states and our agency. Depending on the approach, there are two distinct ways to see self-knowledge as special. If we focus on epistemic access, we may feel that there is something special about the way we come to know or observe things such as our beliefs, attitudes and dispositions. Call this *specialness of access*. By contrast, if we emphasise the role of agency in self-knowledge, we may wonder whether talk of “coming to know” our beliefs or attitudes is misleading. In that case, we may be inclined to see the specialness of self-knowledge in the fact that at least sometimes we seem to create it rather than discover it. Call this *specialness of agency*.

This investigation is not about specialness of access. In what follows, I assume that the way we come to know our mental states is not interestingly different from the way we come to know anything else. More specifically, I assume that all knowledge, including knowledge of our mental states, involves inferences based on the cognitive architecture and previous experience of the knower. Because inference is a type of mental happenings, this means that all knowledge, including knowledge of our mental states, involves a certain type of mental happenings. It just so happens that sometimes these mental happenings have to do with other mental happenings within the same biological organism. When they do, we may talk of knowing our mental states.

If this investigation is not about specialness of access, then the dichotomy set up above suggests that it must be about specialness of agency. That is indeed the case. However, an important qualification is in order. Accounts of self-knowledge that focus on agency tend to be interested in a particular kind of agency, namely rational agency (Gertler 2010). Roughly, they say that I come to know my beliefs and attitudes by deliberating about my reasons for having those beliefs and attitudes. In contrast to that, my point of departure is that the agency-related specialness of self-knowledge goes beyond rational deliberation or, for that matter, mental states, though it can and does concern both. I will use a couple of examples to show what kind of specialness of agency I have in mind.

Suppose I know a fact about something other than myself, say that raccoons have striped tails or that Sweden does not use the euro. For the sake of this demonstration, assume a JTB view of knowledge. Suppose, in other words, that I have two justified true beliefs: one about raccoons and one about the Swedish monetary system. Granted, at some point in the future
both the coloration of raccoon tails and Sweden's currency might change. Then I will need to update my beliefs accordingly. It would, however, be distinctly odd to wonder now whether these true beliefs I currently have about raccoons and Sweden are really about raccoons and Sweden rather than about something else. It would also be odd to wonder if they have ever been about raccoons and Sweden rather than something else. In other words, if I have compelling reasons to hold predicate $x$ true of subject $A$, it does not normally occur to me to question my belief by questioning my identification of subject $A$. Even when I do encounter a reason to doubt my belief that $A(x)$, the doubt typically concerns the predicate rather than the subject, textbook cases of mistaken name attribution notwithstanding.

Learning things about myself seems to be different. Consider two justified true beliefs I currently hold about myself: that I speak English and that I like Icelandic indie music. The relevant evidence is overwhelming. I teach English for a living; I am using English to write this essay; I have memories of speaking the language on countless occasions; I can open my mouth and produce a string of English sentences right now. As for my fondness for Icelandic indie music, my Spotify account features a long playlist with Icelandic indie artists. I distinctly remember both compiling that playlist and listening to it with considerable pleasure for hours on end. What is more, I can turn on that playlist now and experience a characteristic rush of music-induced exhilaration as a track by Valdimar or Ásgeir kicks in.

In other words, the mere act of attaching a linguistic predicate to myself seems to suggest a follow-up question: is that really me we are talking about? It is worth noting that the use of *really* here and later in this section is purely emphatic. The question is “Is it me?” Consequently, no distinction is implied between a “real me” and a me that is “not really me”.

It might also be worth noting that I do not see as philosophically consequential the difference between first-person statements involving the English subject form *I* and first-person statements involving the English object form *me*. The syntactic marking of pronouns as well as its relation to the (presumed) semantic roles of agent and patient vary greatly across languages, and in a number of languages the pronominal morphemes carry no syntactic marking at all.

Finally, it is worth stressing that by “predicate” I always mean linguistic predication, as in “‘Like Icelandic music’ is the predicate in the sentence ‘I like Icelandic music’”. Any other kind of predication (assuming that it cannot be traced back to linguistic predication) is not relevant to my discussion.
sentences and responds enthusiastically to Icelanders strumming guitars. But are those predicates true of *me*?

It is relatively easy to come up with an account on which they are not. To begin with English, it is not my first language. I didn’t know any English until I was ten, and I didn’t become fluent in the language until about 19. Speaking English feels quite different from speaking Russian, my mother tongue. My voice sounds different. Sometimes I catch myself saying things in English that I might not say if I were speaking Russian – and vice versa – simply because they might not occur to me were I drawing on a different set of patterns known as “a language”. Who is to tell whether I remain *myself* when I’m using English? Perhaps my entire English-speaking persona is and has always been a lie, an alien behavioural pattern imposed on me by Anglophone cultural hegemony.

By the same token, my fondness for Icelandic indie music might not be part of who I really am. What if it is largely due to unsavoury sociocultural conditioning? In Western popular imagination, Iceland is one of the “whitest” countries you can find. It is a Scandinavian island that was cut off genetically and culturally from most of the world for hundreds of years. Perhaps my emotional response to Icelandic music is a sophisticated kind of racist knee-jerk reaction, one among many similar reactions that I strive to uncover and distance myself from.

Those were not isolated examples. As soon as I attach a predicate to whatever linguistic sign I use to self-refer, questions such “Is that really me?” or “Is it really part of who I am?” seem to present themselves. Admittedly, for some values they present themselves less readily than for others. For instance, it might take familiarity with gnostic Christianity or Parfit-style thought experiments to see that “Is the human being in the mirror really me?” is not an entirely idle question. With a bit of philosophical probing, however, the *I* of *I PREDICATE* threatens to detach itself even from the most familiar predicates.

Moreover, it seems that this detachment does not even require a rationale of the sort that I sketched above as a way to alienate myself from speaking English and liking Icelandic musicians. The mere act of posing the question (“Is it *me* that speaks English?””, “Is it *me* that likes Icelandic indie music?”) makes it possible to answer the question in the negative – for whatever reason or for no articulated reason in particular. Consequently, no matter what
proposition I come to know about myself, there is a sense in which that knowledge is a matter of choice: I can choose either to distance myself from that proposition or to identify with it until the next time I pose the question “Is it me that p?” Indeed, can is the wrong modal here since I have to make that choice once I have posed the question. This experience of having to choose what I know about myself every time I put my mind to it is what I try to make sense of in this investigation.

1.3 Method

This essay has a “Method” section. This is because the method of my investigation goes somewhat beyond the default, which I will call the Method of Anglophone Philosophy (MAP).

For several decades, the dominant strand within Anglophone academic philosophizing has been what is often called “analytic philosophy”. As noted by Grahame Lock (1999), “post-analytic” might be a more appropriate label. The prefix post- would distinguish this form of philosophical enterprise, tracing its pedigree to Quine’s attack on the analytic-synthetic distinction (Quine 1951), from analytic philosophy proper, which predates Quine. However, the use of the adjective “analytic” to describe anything that involves the MAP is probably too entrenched to deviate from. This is why I will stick to it, albeit with a constant nod to Lock’s observation. The following brief description of analytic philosophy and its default method is based on Crone (2016), Mukerji (2016) and Rathgeb (2019).

Very broadly, analytic philosophy is neither therapeutic nor revelatory, at least not as its primary goal. Instead, it sees itself as part of the larger scientific project to understand the natural world. It is supposed to proceed – some would say “progress” – in lockstep with the natural sciences. The core of its method, i.e. the MAP, is analysis of concepts, as expressed in language. For the purposes of MAP analysis, concepts are implicitly understood in functional terms: they are cognitive tools that we use to organize our experience in order to make predictions and plans. The goal of MAP analysis is to clarify the conditions under which concepts apply, typically by laying bare their connections to other concepts. To that end, the philosopher describes speech acts or scenarios involving the target concept; then she

---

2 This definition of the concept of concept is based on the discussion in Lalumera (2009).
uses her intuition to tease out implications. The scenarios are often imaginary or even unrealistic; the only strict requirement is that they be logically consistent. However, realistic scenarios – including actual observations supplied by empirical sciences – can serve just as well or indeed be preferred. Empirical data can be cited to illustrate or even lend weight to an argument. As a general constraint, conclusions must not directly contradict any well-established scientific findings.

The MAP is the main method of my investigation. I take philosophy to be the practice of creating, combining, clarifying, evaluating and discarding concepts in an attempt to make sense of what is, or what ought to be. As I understand this practice, it requires a constant feedback loop between armchair reflection and empirical research. At the end of the day, “which constructs are permissible” in our efforts to make sense of reality depends on “how the world turns out to be” (Fodor, 1985, p. 153). Whether any given concept applies to “phenomena in the actual world depends on science (broadly speaking) and discovery of the facts” (Churchland, 2017, p. 91). These assumptions are taken for granted in the MAP as it is practised today. Having said that, it is worth stating explicitly in what ways this investigation goes beyond the core of the MAP, i.e. conceptual analysis. At various points, I draw on evidence from the actual use of natural languages as well from empirical findings in psychology, cognitive science and related fields.

One way in which I stray from the MAP particularly far warrants a detailed note. I take it as a given of the human epistemic condition that there is a difference between the observer perspective (roughly, the contemplative third-person viewpoint) and the participant perspective (roughly, the practical first-person viewpoint) on the world. This distinction ultimately goes back to Kant’s *zwei Standpunkte* from which a human being can contemplate herself (GMS, AA 04: 447-453), though I must stress right away that I do not see the two perspectives as representing distinct realities. At different points in my investigation, I discuss the two perspectives and their implications for the target phenomenon in detail. For now, suffice it to say that the MAP, just like the scientific enterprise in general, presupposes the observer perspective on phenomena. My working assumption, however, is that we cannot do justice to the phenomenon of self-knowledge if we only consider it from the observer perspective. This is why I also bring in the participant perspective.
The branch of philosophy most closely associated with the first-person viewpoint and thus the participant perspective is phenomenology. As Katja Crone helpfully puts it, the phenomenological method does not aim to explain phenomena; rather, its goal is to describe how they appear to a conscious subject in a way that is sufficiently systematic and precise to serve as a basis for discussion with other subjects (Crone 2016, p. 11). In other words, to have a meaningful discussion of self-knowledge we need an account of—as Thomas Nagel might put it—what it is like to be a self. For this reason, an important part of my investigation draws on Sartre. I argue that Sartre’s phenomenological account in L'être et le néant (Sartre 1994) captures a key, if not the key, element of what it is like to be a human self in the actual world. It is worth stressing, though, that I embed the phenomenological strand of my essay in an evolutionary account of human selfhood from the observer perspective.

Before wrapping up this section, I should mention two more things, namely my view on the practice of philosophical writing and my reasons for pursuing this project. As regards philosophical writing, I strive not to treat it as transparent, i.e. as if it were a standardised vehicle for communicating certain information. This is because, as noted earlier, philosophy in general and analytic philosophy in particular work with concepts, expressed in language. Language is not transparent; it is a messy system of relations involving facts, speakers, classes of signs, instances of signs, relations among signs, and so on (see e.g. Russell 1959, pp. 132, 145-155). This is as true of modern academic English as it is true of any other language variety, and one need not engage in full-blown post-structuralist deconstruction in order to see that philosophy writers might do well not to pretend otherwise. Philosophical prose is a genre of creative writing, alongside literary prose or journalism, with the added genre-specific expectations of dialectic rigour and heightened self-awareness. In analytic philosophy, the rigour requirement is characteristically emphasised. The “ideal self-image of the analytic philosopher”, as one French observer puts it, “is that of a mathematical logician” (Lecercle 1999, p. 12, my translation). My working assumption, however, is that self-awareness is just as important as rigour. A philosophical text is written at a particular time for a particular audience by a particular person enculturated in a particular society. Most evidently, it is written in a particular language. With all of that in mind, my writing throughout this essay is my best attempt at a compromise. On the one hand, I strive to use appropriate MAP clichés in order to limit misunderstanding and make my arguments easy to trace. On the other hand— and to the extent that I can do it at all in a language other than
Russian – at times I have to go beyond those clichés, both in order to keep my biases and idiosyncrasies clearly visible and to do fruitful conceptual work in the first place.

Finally, my reasons for this investigation are worth mentioning because they inform both my choice of method and its application, as well as my approach to philosophical writing. One of those reasons is suitably theoretical: one day I hope to have a go at a theory of well-informed choice, and this essay lays some philosophical groundwork for that future effort. My other goal, however, can best be described as personal. “What does it mean to know thyself?” is a question I need answered. As Bryan Magee puts it in Confessions of a Philosopher, it is a question “about the situation in which I immediately” find myself (Magee 1999, p. 8). I have to grapple with this question if the human condition – my condition – is to make any sense to me at all.

This concludes the note on method. According to the MAP, a possible next step is to offer provisional definitions of the target phenomenon or its elements, which could then be used, refined or rejected in the analysis. My target phenomenon is self-knowledge. Consequently, I need an account of the self and at least a working definition of knowledge. In the sections that follow, I begin with the self.
Chapter 2. The Self

The goal of this chapter is to define the self. In 2.1 (“Moore’s distinction and no-self views”), I stress the distinction between showing that the term self does not refer to anything real and showing that the term self does not refer to a particular type of thing. I argue that most no-self views are of the latter kind. While they show that various traditional forms of reified selves are untenable, they do not show that self does not have any real referent at all.

In 2.2 (“The self and the singular first-person pronoun”), I assume a close semantic link between the term self and whatever the singular first-person pronoun refers to. I use cross-linguistic evidence to argue that this assumption is unproblematic. Then I consider some of the key rules for using the singular first-person pronoun and discuss the implications of these rules for selfhood. To have a self, I argue, an entity must be capable of conscious self-reference.

My next step towards a definition of the self, which I take in 2.3 (“The Writer from the observer perspective”), is to identify the necessary conditions for the emergence of entities capable of conscious self-reference. I frame this step as a description of the Writer, i.e. the conscious self-referring entity writing this essay, from the observer perspective. The Writer, I conclude, has a self in virtue of being a particular type of biological-psychological-social event cluster: he is a distinct localized unified experiencer-agent that interacts with other agents and self-refers using language. Some of these pre-conditions for selfhood are associated with the notion of a basic pre-reflective self, defended by a number of authors. I reject pre-reflective selfhood as superfluous. Self-consciousness, I argue, is object consciousness turned upon itself, and selfhood is therefore necessarily reflective and complex.

2.4 (“The Writer from the observer perspective”) takes the next step towards a definition of the self by describing the Writer from the participant perspective. The section begins with a clarification of the notion of the participant perspective. Then I draw on Sartre’s phenomenological account in L’être et le néant in an attempt to zoom in on the minimal enduring referent of my singular first-person pronoun. From the participant perspective, I conclude, this referent can best be described as a series of unavoidable future choices, which
I call a Choice-Maker. As part of the discussion in 2.4, I draw two conceptual distinctions. One is the distinction between past choices and future choices. The other is the distinction between free choice and choice as such. Both distinctions, I argue, are crucial if we are to make sense of Sartre’s account and use it to understand selfhood.

Section 2.5 (“Bridging the perspectival gap”) creates a bridgehead for an attempt to cross the gap between the participant perspective and the observer perspective, which I then undertake in the last part of chapter 2. In 2.5, I discuss perspectival translation as such and explain why I expect the notion of a Choice-Maker to be translatable to the language of the observer perspective. I then clarify why rendering the concept of choice into the observer perspective appears to pose a particular challenge. I conclude 2.5 by suggesting why we can hope to rise up to that challenge.

Finally, in 2.6 (“The self as a reflective matrix of maybe”) my two main strands of evidence – from empirical sciences on the one hand and from phenomenology on the other – come together in an evolutionary account of selfhood. I identify the minimal enduring referent of the Writer’s I, described as a Choice-Maker from the participant perspective, with what Baumeister et al. (2018) call a matrix of maybe functioning within a contingency matrix. In other words, I define the self as a reflective mechanism for imagining, evaluating and choosing among alternative futures on the basis of the evidence available to the organism.

2.1 Moore’s distinction and no-self views

In a late essay entitled “Mina filosofiska fördomar”, Konrad Marc-Wogau stresses at some length the importance of a distinction that he attributes to G. E. Moore and therefore calls den mooreska distinktionen (Marc-Wogau 1977). Moore’s distinction, according to Marc-Wogau, is the difference between the reference of a term on the one hand and a particular theory about the reference of that term on the other. Ignoring Moore’s distinction can lead someone to claim that a term does not refer to anything when they ought to be claiming instead that the nature of its reference is not what some people think. Marc-Wogau’s example of such confusion involves the concept of rights. Apparently, Hägerström and his followers were given to statements like “There are no rights”, even though what they were reacting to was a particular understanding of rights, namely that a right was some sort of
supernatural property possessed by people independently of social and political institutions (Marc-Wogau 1977, p. X).

In the case of selves, Moore’s distinction manifests itself as the difference between maintaining that selves do not exist at all (i.e. the term self does not have any reference in the real world) and holding that selves do not exist in a certain way (i.e. a particular theory of the self is not coherent or borne out by empirical evidence). There is a venerable philosophical tradition of denying that selves exist. A thoroughgoing skepticism about the reality of the self was a key tenet of much of the classical Buddhist thought in India (see e.g. Siderits 2011, Gold 2018). In the Western tradition, Hume famously looked in his consciousness for something like a “simple and continued” self only to find a jumble of “particular perception[s] of heat or cold, light or shade, love or hatred, pain or pleasure”, which did not amount to a single entity he could track over time (Hume 1896). Over the last hundred years, the self has been variously dismantled both in so-called continental philosophy and by thinkers of the analytic persuasion. To name just one notable example of the latter, a section title in Derek Parfit’s classic Reasons and Persons unequivocally announces a “liberation from the self” (Parfit 1987, p. 281). More recently, modern brain-imaging techniques and corresponding advances in neuroscience have inspired a number of attacks on the self that combine conceptual arguments with empirical findings, the most relevant of which can be summed up – admittedly somewhat glibly – as a failure to identify a dedicated “self-module” in the brain. A good example of a no-self account informed by neuroscience is Metzinger (2009, 2011). On Metzinger’s view, the self is an illusion, an ephemeral side-effect of “perceiving, interacting, goal-directed” agency that has emerged in the course of evolution (2009, p. 193).

Now comes my first step towards a definition of the self. I assume that many of the arguments against the self put forward by skeptics over the last two millennia are essentially correct. However, in order to take full advantage of their implications we ought to consider them in light of Moore’s distinction. As already noted, I believe that arguments against the self, while often successful and fruitful, do not usually manage – or even aim – to show that the term self does not refer in any meaningful way. Instead, as often as not they target and undermine specific ideas about selfhood. Collectively, the ideas that many self-skeptics react to can be described as an extreme reification of the self: a tendency to treat the self not simply as an object, but as an object that remains essentially unchanged over long periods of
time, including for all eternity. It is the view of selfhood that Anthony Kenny once bluntly called a “piece of philosopher’s nonsense consisting in a misunderstanding of the reflexive pronoun”, “a mysterious metaphysical entity distinct from, but obscurely linked to, the human being who is talking to you” (Kenny 1989, p. 87). This extremely reified self is the true target of modern “anti-selfers” such as Parfit or Metzinger just as it was the target of the Buddhist thinker Vasubandhu in the 5th century C.E.

A popular label for a reified non-materialist view of the self is “a Cartesian Ego” (see e.g. Parfit 1987). With all due respect to the lasting influence of the Méditations, adopting this label as a general term for reified views of selfhood might be somewhat misleading. Not all such views have been as explicitly dualist as Descartes’s thinking substance. Perhaps even more importantly, the tradition of reifying the self predates Descartes by millennia. This is why I will follow William James in using the label pure Ego instead (James 1918).

One straightforward way to define the pure Ego is by stating what it is not. Some of the ideas about what the pure Ego cannot be have been impressively consistent across centuries and civilizations. To stay faithful to their cross-cultural nature, I will sum them up by relating a dialogue between Indra and a deity named Prajapati in one of the Upanishads, as presented in Sharma (1960). Prajapati comes to Indra and asks him to explain the nature of the self. Eventually (i.e. after 32 years of preliminary penance to put Prajapati in the right frame of mind) the two of them arrive at the idea of a pure Ego by ruling out step by step what the self cannot be. It cannot be the body or any of its parts, for the body changes and ultimately perishes, and the true self can neither change nor perish. It cannot be the perceiving subject of a dream, for the dreaming subject feels emotions and pain, and the true self is always at peace. Finally, it cannot be the subject of dreamless sleep, for such a subject has no content of any kind, and the true self cannot be a mere abstraction.

If the self is none of those things, what can it possibly be? According to Indra, the self is something “universal”, “immanent”, “transcendent” and “self-luminous”; it is “the ultimate subject which can never become an object and which is to be necessarily presupposed by all knowledge” (Sharma 1960, pp. 20-21). If we ignore the characteristic Indian insistence that the true self be “universal”, i.e. devoid of any personality, this reification strategy rings a number of bells for anyone who has read the Phaedo or many of the later Western thinkers (Descartes included) that have written Whitehead’s proverbial footnotes on Plato’s
dialogues. The true self is typically assumed to be “entièrement et véritablement distincte” (Descartes 1920, p. 121) from the body; it is essentially unchanging and easy to track over time (including the time after the body’s death); and it is often expected to dwell in a state of contemplative bliss once left to its own devices, i.e. unencumbered by matter. It is no accident that Hume looks for something “simple and continued”, something beyond a motley stream of sensations, when he tries to perceive his own self.

As already noted, it is against this reified pure Ego that no-self advocates from Vasubandhu to Metzinger typically militate. For Vasubandhu, such an enduring self is a conceptual fiction of limited use; it has no reality beyond the reality of “sensory impressions, ideas, and mental events” (Gold 2018). This classical Buddhist analysis of the ultimately “unreal”, i.e. aggregate, nature of the pure Ego bears a striking similarity to many modern no-self views in the West. Here is William James 15 centuries later, arriving at the same conclusion: “the 'Self of selves,' when carefully examined, is found to consist mainly of the collection of … peculiar motions in the head or between the head and throat”; by “motions” James means “the acts of attending, assenting, negating, making an effort”, and other events usually described as “mental” (James 1918, pp. 300-301). Here is Metzinger another century later: the illusion of the self is due to our “hypostatizing a phenomenal individual where there is only an intermittent chain of events, using ‘I’ as a designator where it is only an indicator” (Metzinger 2011, p. 294).

Metzinger goes on to observe that debates about the reality of the self tend to deteriorate into “merely ideological disputes, turning into projection screens for metaphysical desires and hopes”. As examples of “ideological reasons” for the reluctance to embrace the no-self view, he singles out the Western religious doctrine of the soul and “some forms of philosophical phenomenology” (Metzinger 2011, p. 294). This observation strikes me as only too accurate. As I have said earlier, I find arguments against the self both convincing and fruitful. Moreover, the emphasis on the non-existence of the self is understandable, considering the hold that the reified pure Ego seems to have on the so-called common sense. However, that is a rhetorical rather than philosophical reason. In a purely philosophical context, as Marc-Wogau took pains to stress, we ought to respect Moore’s distinction between the reference

---

3 It is possible to argue that James did not subscribe to a no-self view in the same sense as recent thinkers such as Parfit or Metzinger; see e.g. Meini 2012 for such a reading. By quoting James here, I do not mean to label him as a no-seller. The illustration works even if James was mindful of something like Moore’s distinction and was specifically attacking the reified self.
of a term and a particular theory about that reference. Moore’s distinction forces me to
distinguish between the pure Ego, which almost certainly does not refer to any real
phenomenon, and the self, which might so refer. In other words, I reject extreme reification
of the self, along with any dualist ontology typically inherent in that maneuver, be it mind-
body dualism, subject-object dualism or noumenal-phenomenal dualism. However, I also
believe that the first-person pronoun is not always a mere indicator, to use Metzinger’s term.
In the rest of this chapter, I argue that the first-person pronoun does designate something,
and I suggest an evolution-based ontology of the phenomenon it designates.

2.2 The self and the singular first-person pronoun

Stefan Lang opens a monograph on performative self-consciousness by noting that
philosophical discussions of self-consciousness often begin with a look at the singular first-
person pronoun (Lang 2019, p. 7). In that sense, my investigation is typical. For its purposes,
I treat the meaning of the term self as grounded in the use of the first-person pronoun I. To
be more precise, I assume that whatever parts of the world the first-person pronoun picks out
in various speech acts form the basis of our intuitions about what the noun self picks out
when used in philosophical or scientific discourse. I see this assumption as unproblematic
for a cross-linguistic reason. The reason might need some elucidation in the form of a
digression from the core of the MAP. The following reflection on what happens in natural
languages is meant both to solidify the link between selfhood and linguistic usage and to
convince the reader that, however wrongheaded my view of the self might be, it is not based
on “a misunderstanding of the reflexive pronoun”, to use Kenny’s words quoted earlier.

First, consider languages such as English, German or French. In those languages, the nouns
typically used in contemporary scholarly prose about selfhood derive from reflexive
pronouns. The reflexive pronouns in question may co-refer with subjects of all persons (the
English myself, yourself, itself etc. or the German selbst, as in the Duden’s examples du
hast es selbst gesagt and sie muss alles selbst machen); or they may only co-refer with third-
person subjects (the French soi). This linguistic accident, trivial in and of itself, may obscure
the close semantic kinship between the noun (the self, das Selbst, le soi), as it is used in
scholarly parlance, and the first-person pronoun, as it is used by speakers across the board to
refer to themselves. Given the increasing dominance of English in philosophy and its near-

4 See https://www.duden.de/rechtschreibung/selbst.
monopoly in published science, the etymological independence of the philosophical noun *self* from the mundane indexical *I* may be particularly worth noting.

Luckily, other languages have often chosen to nominalize the first-person pronoun itself for use in philosophical debates and cognitive science.\(^5\) Consider these examples from Italian, Russian and Swedish:

**Italian:** Thomas Metzinger afferma che l’io non esiste, nel senso che esso non è altro che il contenuto di un particolare modello costruito dal cervello, il cosidetto «modello fenomenico di sé»… “Thomas Metzinger claims that the self [lit. the I] does not exist, in the sense that it is nothing but the content of a particular model created by the brain, the so-called ‘phenomenal self-model’ [lit. model of itself]…” (Marraffa & Paternoster 2013, p. 62, my translation). Note how Italian employs both a noun derived from the first-person pronoun (*io*) and a nominalized third-person reflexive pronoun (*sé*).

**Russian:** Тожество я, определяющее сущность эмпирического я, не является, следовательно, противоречивым понятием… “The sameness of the self [lit. sameness of I, *ya*], which determines the essence of the empirical self [lit. empirical I, *ya*], is therefore not an incoherent notion…” (Shpet 2020, p. 494, my translation).

**Swedish:** Hela Merleau-Pontys filosofiska strävan gick ut på att upphäva dualismen mellan medvetandet och materien, jaget och världen… “In all of his philosophical work, Merleau-Ponty aimed to overcome the dualism of consciousness and matter, the self [lit. the I] and the world…” (Nordin 2013, p. 570, my translation).

We can add the Latin term *ego*, used in a number of languages, to this list. The close affinity between the nouns *ya*, *io*, *jag* and *ego* on the one hand and the phonologically identical subject forms of the first-person pronouns on the other is evident. To be clear, these nouns are not analogous to the occasional nominalization of *I* in English (as in “Who is this other I that doubts my every step?”). They can and do occur in many contexts where English would talk of *selves*. In fact, as we have seen, English-speaking authors have also used the Latin *Ego* as a synonym for *self*.

---

\(^5\) Indeed, even the modern German use of *das Selbst* is a calque from English, according to the Duden; see https://www.duden.de/rechtschreibung/Selbst. In the past, German-speaking thinkers were apparently happy to nominalize the first-person pronoun. Freud’s *das Ich* is one well-known example.
This is why I consider it unproblematic, if not downright trivial, to treat the meaning of the noun *self* as grounded in the use of the first-person pronoun across a variety of speech acts. This is not to deny that in specialized literature the term *self* may acquire technical senses that are far removed from whatever ideas the layperson has about themselves. Indeed, in the course of this investigation I might end up defining a self that is rarely, if ever, picked out in that exact form by the first-person pronoun in everyday speech. However, my key point is that the discussion itself is an attempt to understand something crucial that always lurks behind the reference of *ya, I, jag* etc. whenever I use them to refer to myself. It is an attempt, in other words, to zoom in on what we may call the minimal enduring referent of the singular first-person pronoun in any first-person statement that involves self-reference: the referent that I can no longer choose *not* to refer to at least as part of a larger referent once I realise that I can refer to it.

For this reason, it is worth recalling some key rules for using the singular first-person pronoun in non-reported speech. What follows is a version of the list from Lang (2019, p. 7), modified in particular to exclude any mention of persons:

(a) *I* always points at the speaker that uses it.
(b) Whenever *I* is used by an existing speaker, it points at something that exists.
(c) Whenever a speaker uses *I*, she knows that *I* points at her.
(d) Whenever a speaker uses *I*, she knows that she uses it.

(a) means that an act of self-reference, which is what using *I* is, is not an act of self-reference unless the sign used points at the user. In (c) and (d), I take “knows” to refer to knowledge implicit in the speech act rather than a conscious event of meta-awareness about one’s language use, though such meta-awareness may certainly accompany a speech act involving a singular first-person pronoun, e.g. when the speaker is using a language she is not quite fluent in. With that clarified, here is how these rules are relevant to my discussion of the self.

First off, (b) states that in order to use *I* correctly the user and thereby referent of *I*, whatever it may be, must exist in the actual world. It is, in other words, must be unequivocally real.

---

6"1. Der Ausdruck *ich* referiert auf diejenige Person, welche diesen Ausdruck verwendet. 2. Im Fall der Verwendung des Ausdrucks *ich* existiert der Referent dieses Ausdrucks. 3. Im Fall der Verwendung des Ausdrucks *ich* weiß der Sprecher, dass dieser Ausdruck auf ihn selbst referiert. 4. Im Fall der Verwendung des Ausdrucks *ich* weiß der Sprecher, dass er selbst diesen Ausdruck verwendet."
As I have argued, even strong arguments against the self do not contradict (b). At most, what they show is that some of the referents people take I to point at are not real. If we take I to pick out a pure Ego that endures essentially unchanged over long stretches of time, then I fails to pick out anything real. However, every token of I in non-reported speech does pick out something real that does not fit the definition of a pure Ego. This, to re-state (b), is a pragmatic prerequisite for a correct use of the first-person pronoun.

Next, both (c) and (d) suggest that the speaker, whatever she is, must have certain properties. These properties are worth spelling out. Firstly, the speaker must use a sign system (e.g. English, Swedish Sign Language, or the system used by Kanzi the bonobo) that allows her to refer to herself. Secondly, she must have an awareness, or be conscious, of using that system. Finally, she must have an awareness, or be conscious, of being something distinct from the rest of the world – something that her sign system allows her to pick out by using a sign such as I.

At this point one might counter that (c) or (d) do not say anything about having awareness or consciousness. Do speakers really need to be conscious in order to know that they use I to refer to themselves? After all, even a basic customer service bot seems to use the first-person pronoun quite well without being conscious of anything at all. Should there not be a sense in which such a bot knows that it uses a linguistic sign to refer to itself? Even if the answer is yes, I assume that such a sense of know is non-literal. In a later section, I clarify in detail the view of knowledge that I adopt for the purpose of this investigation. However, it is worth stating early on that I take knowledge to involve conscious information processing. Bots and other non-conscious I-users have information about using the first-person to refer to themselves; they do not know that they do it.

To be clear, my chief concern here is not that we should revise (c) and (d) in light of a distinction between knowledge (a property of conscious things) and information (a property that does not require consciousness). Rather, my point is this. It may well be that discussions of philosophical zombies and experience of interaction with non-conscious I-users such as online bots affect our intuitions about the use of the singular first-person pronoun. However, whatever changes they bring have little or no bearing on the issue of selfhood, as it has been debated since the Axial Age and as I tackle it in this essay. To put this in terms of prototype semantics (Rosch 1978), I assume that prototypical instances of I-users are conscious; any
non-conscious I-users lie on the periphery of the category. For the rest of this essay, I ignore non-conscious I-users and focus exclusively on conscious self-reference.

2.3 The Writer from the observer perspective

The I-user I am most intimately familiar with is me. This is why I frame this and the following sections as a description of the entity writing this essay, henceforth the Writer. An additional bonus of keeping the discussion focused on the Writer is that the only participant perspective I can adopt is my own. This becomes helpful in the next section, where I consider the Writer from the participant perspective. This section, for its part, describes the Writer from the observer perspective by placing him in his evolutionary context. Since the observer perspective is the default perspective of science and the MAP, this section offers no explicit definition of what it means to adopt it. My hope is that whatever may be unclear about the observer perspective becomes clear in the next section by contrast with the participant perspective, which I do define at some length.

From the viewpoint of biology, the Writer is an individual of the animal species Homo sapiens. Understood dynamically and diachronically, he is a “self-organizing biological event”, “roughly the sum of the metabolic activities” that his parts are involved in (Olson 2007, pp. 28, 136). Roughly, this biological event began in the summer of 1978 in the north-west of the Soviet Union, i.e. at the spatiotemporal point of conception, and it will end upon the cessation of its metabolic activities at an as yet unspecified spatiotemporal point in the future (hopefully, a suitably distant one). The fundamental identification of the Writer with a biological organism (Olson 1997) is illuminating for many reasons. One of them is the emphasis it puts on the Writer’s place in biological evolution. The fact that the Writer, whatever else he might be, is an outcome of evolution becomes important later in this section.

First, though, I need to note that for many other reasons defining the Writer as a cluster of metabolic activities is not enough. Once humans are born, they start interacting with other humans. Other types of activities come into play and begin to define them. After his birth, the Writer gradually became a “complicated biological-psychological-social” forensic unit.

---

7 I use the masculine pronoun to refer to the Writer out of linguistic habit even though the form it might be more appropriate in this context.
(Schechtman 2014, p. 198). In other words, in his current form the Writer is a causally unified cluster of physiological, neurological and sociocultural events that has a particular moral status. A cluster of this kind is what I will have in mind whenever I use the term person (with the qualification that Schechtman’s term person life is more accurate than person because of its diachronic aspect). It is worth stressing that I do not intend to use person and self interchangeably. The notion of a person – as a forensic unit of the sort described above – strikes me as relatively unproblematic and distinct from the self. This is not to deny that in specific cases it can be difficult to decide whether a human animal has enough psychological or social properties to qualify as a person. Foetuses and vegetative state patients are two textbook cases in point. Moreover, the typical list of person properties may well change as our technology develops, allowing scenarios that are now only possible in science fiction. Regardless of how we update the concept of person, however, I believe that it is unlikely to merge with the concept of self, as I define it below. While many of the conditions necessary for personhood may be the same as for selfhood, there also seems to be a crucial difference: we do not need to bring in the participant perspective to define or track a person as a forensic unit. The observer perspective appears sufficient for that purpose. That is the view regarding persons that I assume from now on.

Let us return to the Writer. At some point, as he was acquiring features associated with personhood, the Writer became self-aware. Because I intend to argue that having a self presupposes robust self-awareness, I can also describe this event as follows: the Writer developed what it takes to be a self. For an account of how that happened, i.e. for an ontogeny of the human self, I largely rely on Marraffa and Paternoster (2013). They offer a plausible synthesis of conceptual analysis with findings in cognitive science and developmental psychology.

To start with their conclusion, Marraffa and Paternoster see the self as a complex psychological function (2013, pp. 71, 73-74, 112-113). Similar to other complex functions, the self is not localized in any particular neurological structure. The brain’s self-organizing activity is distributed; it does not converge on a “self module” (see also Singer 2008). The self is constructed and maintained by various subsystems involved in action, memory, socialization and language. This suggests that the self is absent in non-linguistic animals and does not fully emerge in human children at least until around year 3.
A key step towards this view is to make a distinction between consciousness and self-consciousness. By “consciousness”, Marraffa & Paternoster mean object consciousness, i.e. phenomenal representations that have as their intentional content various features of the world. Self-consciousness, by their lights, is a specific case of object consciousness. The self depends on object consciousness conceptually, phylogenetically and ontogenetically. In other words, consciousness comes first both in evolution and in a particular human life; it is a property that the Writer shares with very young children and non-human animals.

This way to present the relation between consciousness and self-consciousness is controversial. A competing view is that the dependence goes the other way or that the two kinds of consciousness are co-dependent. Either way, a number of authors have argued that consciousness always comes hand in hand with a basic, pre-reflective form of self-consciousness. As long as a pre-linguistic child or a non-human animal feels anything at all, so the thinking goes, its feelings must have a certain “mineness”, or “first-person givenness”, to them (Zahavi 2014). Other features typically associated with a pre-reflective self are a sense of owning one’s inner life, a sense of agency, and a first-person perspective (Crone 2016; Marraffa & Paternoster 2013). To what extent these features can be distinguished from each other is a matter of some debate, and there is the added difficulty of conceptual mapping across languages (e.g. which English label matches the German präreflexive Selbstvertrautheit?). Regardless of the details though, the general view adopted by advocates of the pre-reflective self is clear: one or a combination of such features must be immanent in any conscious experience. What it is like to be a conscious thing, the argument goes, is necessarily what it is like to be a conscious thing distinct from the rest of the world and acting upon it from a particular point in spacetime. Therefore, conscious experience has to “belong” to a distinct, localized and unified experiencer-agent. The sense of being that experiencer-agent is the pre-reflective self.

I am not in a position to deny that in some contexts the notion of a pre-reflective self can do useful theoretical work. For example, some form of pre-reflective self is widely assumed in clinical psychology. In clinical research, the term *self-disorder* refers to a family of “qualitative anomalies of subjective experience” (Cermolacce et al. 2007, p. 705). In the context of such disorders, the *self* serves as an umbrella term for whatever it is that malfunctions when people experience alien limbs, thought insertion, failure of self-recognition, as well as other disturbances of body schema, agency and personal identity.
Interestingly, the working definition of “the normal clinical self”, if it is made explicit at all, might vary from study to study, but the general idea seems to be close to a combination of the description I have used above with some features of personhood. In other words, the clinical self appears to be a distinct, localized and unified experiencer-agent that is aware of persisting and remaining relatively stable over time. A breakdown of any element in this cluster of properties can be classified and treated as a self-disorder.

However, the context of this section is that of the Writer’s ontology – or, since the Writer is human, of human ontology. As regards human ontology and related areas of philosophy, this investigation assumes the point of view defended by Marraffa and Paternoster (2013): that the pre-reflective self, while not incoherent, appears to be superfluous. In other words, whatever theoretical work it may do in philosophy can be done more parsimoniously by consciousness as such. As becomes clear in later sections, this rejection of the pre-reflective self is important for my overall argument. Unfortunately, a sustained critique of the pre-reflective self would go well beyond the scope of this essay. I will, however, briefly sketch two immediate reservations which, together with other evidence, lead me to reject the notion of a pre-reflective self. The purpose of this sketch is to show that my reasons for rejecting pre-reflective self-consciousness are independent from the goal of this investigation.

The first reservation about the pre-reflective self that I will mention here is based on the evolutionary origin of consciousness. As we have seen, one fact about the Writer’s ontology is that at any point in his existence he is a biological event. In biology, as the famous essay title has it, nothing “makes sense except in the light of evolution” (Dobzhansky 1973). Keeping that in mind, I side with Marraffa and Paternoster (2013): in an evolutionary account of the Homo sapiens, self-consciousness must be secondary to self-less object consciousness, or consciousness as such.

Here is a general argument to that effect, inspired by several specific objections in Marraffa and Paternoster (2013). Firstly, note that the only form of consciousness we are familiar with is found in biological organisms. In other words, being conscious is something that is done by an organism; we have no evidence that it is ever done by something that is not an organism. Secondly, recall that a biological organism, whether conscious or not, is a distinct, localized and unified agent. Thirdly, add consciousness to that set of features. Now you have a distinct, localized and unified experriencer-agent. In other words, you have what you get
when you add up features said to amount to a pre-reflective self: you have agency, proprioception, a unified body schema, a sense of body-world boundary, a spatiotemporal centredness of perception, and so on.

To be sure, all of these features of consciousness remain non-reflective until the organism develops an ability to reflect on them. Equally evidently, all of them belong to the organism. However – and this is the step that casts doubt on the need for a pre-reflective self – what else could they possibly belong to? At the risk of belabouring the point, recall that consciousness is a function of biological organisms. Any instance of consciousness necessarily belongs to a particular biological organism. What is more, any instance of consciousness likely belongs to a biological organism in the same sense that any instance of digestion, ovulation or blood circulation belongs to a biological organism. At any rate, we have little reason to assume otherwise. To be sure, some of us might have a strong intuition to the contrary: that conscious events must belong to the organism in some other, special, way. However, this intuition may well be due to a common, if not always voiced, assumption that consciousness is fundamentally more mysterious than other biological functions. Even if, however, we grant that consciousness is harder to reduce to a physical description than digestion or blood circulation, it does not follow that it is an elusive, essentially non-spatiotemporal feature that has to be anchored to an organism in some special, fundamentally mysterious way.

In other words, if we wish to defend a pre-reflective self within an evolutionary framework, there is no obvious reason why we should stop at conscious features. Every other part of the biological event we call an organism can go in just as well. Glasgow (2017) provides an example of such admirable consistency. His view of the minimal self incorporates in one fell swoop everything involved in “self-maintenance, self-reproduction and self-containment” (p. 18). To be sure, making the reference of self co-extensive with that of living organism in this way might seem superfluous. It is hard to see, however, why it would be more superfluous than postulating a pre-reflective self made up solely of conscious phenomena. Both conscious and unconscious features of a biological organism are pre-reflectively involved in keeping the organism alive and well. If we do not need to posit a pre-reflective self to explain blood circulation or immune responses, why do we need a pre-reflective self to explain proprioception?
To conclude my first reservation about the pre-reflective self, we seem to have little reason to postulate pre-reflective self-consciousness in order to lay a phylogenetic or ontogenetic foundation for mature, reflective self-consciousness. Any animal, including any human animal, is “a singular organismic being, physiologically distinct from other bodily beings” (Kyselo 2016) and acting out “a fundamental biological imperative to maintain physiological integrity: to stay alive” (Seth & Tsakiris 2018). These features of the animal condition appear sufficient to explain what it means for experiencer-agents to be localized, distinct and unified.

My other immediate reservation about the idea of a pre-reflective self comes from phenomenology. To begin with, note that I avoided terms such as “mineness” or “first-person givenness” when I listed candidate features of pre-reflective self-consciousness while discussing my first reservation. The reason for that is simple: I suspect that such terms are artefacts of linguistically mediated philosophical self-reflection. The mere fact of allowing them in a description of pre-reflective consciousness begs the question in favour of defenders of the pre-reflective self.

I do not wish to deny that, when I put my mind to it, I discern – or at least easily convince myself that I discern – some aspect of what other authors presumably mean by a sense of “mineness”. This mineness seems to accompany my experience of my body and its behaviour. I assume that it is the sense of having a properly functioning, “undisturbed” body schema as well as properly functioning mechanisms of body-recognition, intentional action, and other neuronal routines that add up to a distinct, localized, unified experiencer-agent. Like many people, sometimes I get a brief glimpse of what it might feel like when this default mineness subsides. Every now and then, a sudden feeling of baffled self-alienation overcomes me, and for a few seconds I stare at my hands, or at my face in a mirror, wondering what on earth these particular hands and this particular face have to do with me.

However, I must confess that I have no idea what it is like to experience such mineness as a baby or an animal, without having internalized language or a habit of philosophical self-reflection. My phenomenological data are hopelessly contaminated by years of being a human person in general and a philosophically inclined Russian Westerner in particular. Whenever I try to focus on the mineness of my experiences, I cannot help but do that as a language-user brought up and educated in a culture that is, not to put too fine a point on it,
obsessed with self-reflection. Trying to distil some pre-reflective phenomenal mineness out of my adult human reflections on the nature of selfhood strikes me as akin to an attempt to see what the world looks like when no one is looking.

It is possible that other authors have powers of non-reflective reflection that I lack. However, considering my other doubts about pre-reflective selfhood, I find that somewhat unlikely. Instead, as noted earlier, I agree with Marraffa and Paternoster (2013, especially pp. 102-105): postulating a pre-reflective self is unnecessary. Non-reflective organisms, such as babies or non-human animals, probably perceive external and internal stimuli by means of the same sort of consciousness, namely object consciousness. Subject consciousness, which we observe in ourselves, is simply object consciousness “turned upon itself”: it is an instance of consciousness whose object is one or more instances of consciousness within the same biological organism. To put this in terms of neuroscience, conscious experience “appears to involve a cognitive process that monitors neuronal activation patterns irrespective of whether these result from sensory input or are internally generated” (Singer 2007, p. 605). This investigation therefore assumes that all subject consciousness, or self-consciousness, is fundamentally reflective. As such, it is a relatively late composite “product of the interaction of a number of cognitive subsystems that perform functions linked to language, socialization, and memory” (Marraffa & Paternoster 2013, p. 113, my translation).

To sum up the discussion so far, I have outlined some of the key conditions for the emergence of a human self. I have argued that the Writer has a self in virtue of being a particular kind of biological-psychological-social event cluster. That is to say, the Writer has a self because he is a distinct, localized, unified experiencer-agent that interacts with other such agents and has a capacity for self-reference involving language. As we have seen, several of these features are often grouped together under the umbrella of a pre-reflective self. Particularly in clinical research, there is a feeling that such a self is “a necessary foundation for the articulation of a richer or sophisticated, reflective, language-bound” selfhood (Parnas & Sass 2011, p. 525, emphasis in the original). Without questioning the usefulness of the notion of a pre-reflective self in clinical research, I briefly outlined my main reservations about postulating such a self in human ontology. Humans are biological organisms that have evolved to maintain a complex form of homeostasis; that fact alone, I have argued, is sufficient to account for the features attributed to a pre-reflective self.
Moreover, I fail to discern phenomenological evidence that might support the idea of pre-reflective self-consciousness. I therefore conclude that my initial intuition, based on the semantic kinship between the self and the singular first-person pronoun, is correct: whatever the self is, it emerges when the organism develops an ability to direct conscious reflection at itself. In other words, the “sophisticated, reflective, language-bound” kind of selfhood is not just the kind we are probably best equipped to look for; it is the only kind we have reason to look for.

So far, the discussion has stayed within the observer perspective, i.e. the default perspective of science and analytic philosophy. Even the single brief foray into phenomenology was framed as a third-person observation, with a cognizing subject (“I”) taking stock of the phenomenal objects available to it in a detached, theoretical fashion. In order to arrive at an adequate definition of the self, however, we need to bring in the participant perspective. This is the task of the next section.

2.4 The Writer from the participant perspective

The last section described the Writer as a human person, a complex event cluster brought about by biological and cultural evolution. It identified several properties of the Writer, and of human persons in general, that facilitate conscious self-reference and thus lay the foundation of a self while not amounting to a self. For a self to emerge, I argued, an event cluster needs to be sufficiently distinct, localized, unified, conscious, active, endowed with language, and capable of reflecting on its own properties.

The goal of this section is to use phenomenological evidence in an attempt to zoom in on the remaining ingredient of selfhood: what I called the minimal enduring referent of the first-person singular pronoun. This referent is what I cannot choose not to refer to at least as part of a larger referent once I realise that I can choose to refer to it by using the first-person singular pronoun. Because this referent is closely linked to choice, zooming in on it requires a change in perspective. As I show in the sections that follow (particularly in 2.5), the notion of choice only makes sense when we look at the world as participants rather than observers.

Up to this point, I have considered the Writer from the observer perspective, i.e. the default perspective of science. This perspective treats the Writer as a fully integrated part of a
causally closed world and thus in the domain of theoretical models that we use to describe any other part of reality. It is perhaps worth noting that the observer perspective does not presuppose a reductionist view of scientific explanation, on which any valid theoretical model must be fully translatable into a model involving the most fundamental theoretical units, such as particles or fields. This investigation therefore assumes that, to capture certain features of the world, we might need concepts and models that do not reduce to the language of fundamental physics. For instance, in order to explain different features of the event cluster I call “the Writer” we need theoretical models from evolutionary biology, psychology, sociology, and so on. Many of these models shed light on important aspects of reality while resisting a straightforward translation to the vocabulary of any other model.\textsuperscript{8} All of them, however, have in common the observer perspective on what they describe. In other words, they all embed the Writer within larger event patterns in order to make explanatory claims (roughly of the form “x causes y”) and predictions based on those claims (roughly of the form “y is likely to follow x”, “y will happen given x and z” etc.).

However, experiencer-agents endowed with language and capable of reflecting on their own properties appear to have two distinct perspectives on reality. As I stated in the section on method, my working assumption is that a satisfactory definition of the self must draw on both of them. This is why this section describes the Writer from the participant perspective. Consequently, the description switches from third to first person. Let me stress that this linguistic manoeuvre per se does not constitute a change in perspective. It is possible to maintain the observer perspective while talking about yourself in any person; I did just that in the last section when I was discussing alleged phenomenological evidence for a pre-reflective self. By the same token, it is possible to discuss the participant perspective without any first-person forms. Having said that, in pragmatic terms the participant perspective is more readily associated with first-person forms. Hence the switch.

Roughly, the difference between the two perspectives is the difference between treating something as an event or as a possible course of action.\textsuperscript{9} An event is anything that is causally embedded in a world, such as the actual world. A possible course of action is

\textsuperscript{8} For a discussion of non-reductionist naturalism, see e.g. Putnam 2012 and 2014. For a similar view, called integrative naturalism and applied specifically to the problem of perspective, see Sturma 2015.

\textsuperscript{9} In other words, it is a pragmatic difference; it is not an epistemic distinction between observing an external reality and experiencing my own qualia. On my view of the difference between the observer and participant perspectives, observing my own qualia, which I take to be part of the world on a par with anything else that exists, assumes the same perspective as observing any other part of the world.
anything that I, i.e. a conscious agent, can do at some point in the future.\(^{10}\) The difference, I should stress, is one of description. As noted in 1.2 above, it is a feature of the human epistemic condition rather than the world’s ontology. All my future actions are also events and should often be described as such. Yet it is a difference I cannot overcome or ignore as long as I remain a conscious agent. When I look at the world as an observer, I have plenty of use for explanations of the form “x causes y” or predictions of the form “y is likely to happen”. Characteristically, however, they only seem to be useful when they do not refer to my own actions. Somehow, they never seem satisfactory when I consider what I can do in the future.

To be sure, I can easily make statements of the form “x causes me to do y” or “(I predict that) I am likely to do y”. In other words, because I am a part of the world, I can refer to a pattern I am causally embedded in. For instance, I can say, “A tap on my knee causes me to twitch my leg” and “I predict with a high degree of certainty that I will travel abroad for my next holiday (because that is what Swedes with my level of disposable income tend to do)”. Clearly, though, these utterances only make sense if in both cases I renounce my agency, i.e. if I place my leg twitching and my trip abroad outside the purview of my choice. In the knee-tapping case, that is easy to do since I do not think of the resulting twitch as a matter of choice to begin with. In the foreign travel case, on the other hand, renouncing my agency is much trickier, if not downright impossible. It seems that I cannot simply make a prediction about my going abroad and then see how things work out. Until the trip actually happens, there is always a sense in which I am choosing to go abroad for my next holiday. Every single moment while the trip is still in the future, I can choose not to go. Even if I choose to “go with the flow” – that is, in my case, if I resign myself to whatever destination my partner picks for our upcoming vacation – I will have to reaffirm my choice to resign myself every time I happen to think about it, and this state of relentless choice-making will go on until the holiday finally begins.

It appears then that the participant perspective involves an awareness of inescapable imminent decisions; an awareness of a range of options that you are apparently forced to choose from. A philosophical classic that treats such inescapable choice as a fundamental

---

\(^{10}\) Because I assume that mental events are part of the world on a par with any other events, my understanding of action is correspondingly inclusive: any mental event in a biological organism that is amenable to the organism’s agency can be viewed as a potential course of action by that organism. Put simply, anything that an organism does on purpose, including whatever it mentalizes on purpose, qualifies as action.
feature of human existence is Sartre’s *L’être et le néant* (Sartre 1994). My next step towards a definition of the self is to draw on Sartre’s account of what it is like to be a self-aware human person, or a *réalité-humaine*, to use his term.

First, though, there is something I need to make clear. Sartre defines his 800-page work as an exercise in phenomenological ontology. Consequently, he frames his conclusions in ontological and metaontological terms: one of his aims is to establish, at the highest level of abstraction, what is and what it means for what is to be. I am happy to share some of Sartre’s ontological commitments. In particular, I assume, as he does, that the non-conscious, non-human mode of being is primary (Sartre 1994, p. 811). However, for the purpose of my investigation Sartre’s ontological claims are mostly irrelevant. Whatever this essay borrows from Sartre should be seen as a helpful description of phenomenological data, not as a claim about the structure of reality at the highest level of abstraction. Consequently, what follows is not an attempt at an exegesis of *L’être et le néant*. My subject is not history of philosophy, and my goal is not to reconstruct Sartre’s thinking as faithfully as possible. Instead, I use his thinking as heuristic scaffolding for my own argument.

With that clarified, let us see what picture of the human condition can emerge from *L’être et le néant*. According to Sartre, when a *réalité-humaine* looks at the world and itself in it hard enough, it comes to realize that there are two distinct modes, or domains (*régions*), of being. Some being is being *en soi*: it is, and it is what it is. Being *en soi* has no room for otherness (*altérité*); the question whether it could be something else – i.e. if it could exist in some other way than it actually is – is meaningless. The observation that being is what it is sounds like a trivial analytic statement, to the point where someone less charitably inclined might call it an unenlightening tautology. Sartre agrees that it can be an analytic statement, but only as long as it describes a world in which there is no consciousness (p. 37).

Consciousness creates a kind of gap (*un trou*) in the uninterrupted fullness of being *en soi* (p. 809) because, instead of being what it is, consciousness “is to be what it is” (“*a à être ce qu’elle est*”, p. 37, emphasis in the original). Sartre’s reason for this somewhat obscure formulation is relatively straightforward: he assumes that all consciousness has intentional content that is distinct from consciousness itself. In other words, all consciousness “transcends itself in order to capture an object” (p. 19); the whole *raison d’être* (pun

---

11 All translations of direct quotes are mine, as are the interpretations that I supply for Sartre’s key terms. The terms themselves are left untranslated whenever possible. All page references are to the 1994 French edition.
intended) of any conscious phenomenon is to be something other than itself. This constant outreach towards something beyond itself – this permanent non-coincidence with itself – forms the basis of what Sartre calls the mode of being *pour soi*.

Throughout *L’être et le néant*, Sartre often appears to use the noun *le pour-soi* as another name for *la réalité-humaine* or *l’être humain*. At least some passages suggest that in such cases *le pour-soi* can be read as a kind of synecdoche rather than a synonym. For instance, Sartre describes the body as something that “belongs to the totality we call *réalité-humaine* as one of its structures” (p. 314), which to me strongly suggests that a *réalité-humaine* is a complex entity akin to Schechtman’s person life, discussed earlier. On this reading, a *pour-soi* would be a part of a *réalité-humaine*, namely the conscious part, which has a mode of being distinct from whatever exists *en soi*, including the other structures of the *réalité-humaine* in question.

Let us now bring back the Writer, which is to say me, in order to tell the rest of the story. So far, Sartre’s account has brought up two things I can refer to when I use the first-person pronoun. In some cases, my use of *I* refers to me the human person, complete with my body and any other relevant structures. In other cases, however, my *I* picks out me the *pour-soi*, i.e. the consciousness that is part of me the human person (“I was born in the wrong body to the wrong parents in the wrong place at the wrong time”, to give a particularly compelling example). Let us stick with this second referent. What features do I have as a *pour-soi*?

According to Sartre, that is not the right question to ask, given the mode of my being. Recall that, as a *pour-soi*, I permanently transcend myself: I am in “the mode of not being what I am and of being what I am not” (p. 377). Like any consciousness, the consciousness that is me is forever directed beyond itself; by definition, it is always consciousness of something other than itself. As a result, even though my consciousness is constantly full of phenomena – my consciousness is phenomena – I do not coincide with any of those phenomena; they are simply a factor of the interaction between me and the world. They are therefore not me, and I am not them. In other words, whatever features I happen to have are, in an important sense, not *mine*.

Again, much as this might sound like the punchline of a Zen kōan, impervious to analytic treatment, Sartre’s no-own-features account of human consciousness builds on two straightforward assumptions. As we have already seen, one of them is that all consciousness
is object consciousness; the other is that consciousness, whatever it might be, is not an autonomous substance (pp. 810-811). In fact, for Sartre it is not a substance at all. By his lights, I am definitely not a thing; I am a process, i.e. something that is happening to the en-soi right here and now (“[Le pour-soi] est été par l’en-soi”, p. 810, emphasis in the original). If we also allow that Sartre’s view, like so many others (see 2.1), is in part a reaction to extreme reification of the self, then his insistence that the core part of human nature has no fixed features of its own no longer sounds especially paradoxical.

More importantly, Sartre’s picture of a locus of consciousness that is none of the things it appears to be is instantly recognizable. Can I identify any content of my consciousness as a fixed feature of me? Whenever I put my mind to it, I do not think I can. Every single phenomenal object that I zoom in on turns out to be a feature of what Sartre calls le donné, i.e. what is given. Le donné is the part of the world that interacts with me – or, to stay more faithful to Sartre’s language, the part of the world that I, the pour-soi, am to be. It includes features such as my body, my location in spacetime, my place in society, my relations to others, and my past (pp. 648-726). In fact, it even seems to include my desires: from transient whims in the here and now to long-term dispositions based on long-held interests and convictions. To be sure, I cannot conceive of not having any of those features at all, but I can easily imagine not having the features that I happen to have. If I were writing this in a language that distinguishes between alienable and inalienable possession, I might say that, as long as the subject is “I”, alienable possession markers should apply to everything.

A brief thought experiment can help to clarify this point. Suppose that the age of fully automated luxury communism or some even more ultimate stage of human development (see e.g. Lem 2014) has arrived, and I am lucky enough to still be alive and not irreversibly senile. All the abundance and opportunity provided by a technological utopia are at my disposal. For all practical purposes, my donné – my that-which-is-given – is no longer all that given because I can choose most of the donné I have to deal with. That is to say, I can live wherever I want; I can modify my body in any way I want; I can delete memories from my enhanced brain and upload new ones; I can quickly form new habits and shape my first-order desires to fit my higher-order preferences. On top of all that, I can hook myself up to customized virtual universes, which feel fully real, and role-play anyone – from Hypatia and Jesus to Bilbo Baggins and Anna Karenina – for years of subjectively experienced time on end.
With so many easy choices and a sufficiently long lifespan, would there be phenomenal content that I would always see as inalienably mine? That seems unlikely. To be sure, right now there are plenty of relatively stable features that I consider an inalienable part of myself. They range from certain features of my physical appearance (an average height, a general outline of my face etc.) to my skills and memories (an ability to play the guitar, cultivated native Russian, recollections of growing up working class in a Soviet mining town etc.) to my relationships with my partner and my closest friends. Right now, I feel that replacing any one of those features would be a loss. Replacing all of them feels hardly distinguishable from dying. I strongly suspect, however, that I would not feel the same way after a decade or two in a transhumanist utopia described above. Sooner or later, I would choose to shed most or all of those features at least for a while and possibly for the rest of my life. I suspect this for a simple reason: like most humans, I have undergone comparable transformations even in my actual, highly constrained, existence. Some of the features that seemed profoundly mine to my 13-year-old self or to my 23-year-old self now strike me as just as profoundly alien.

It is also worth noting here that I am privileged to feel fine about most of the features I now identify with. Human pour-sois who are in the grip of a so-called identity crisis or otherwise unhappy with themselves would probably jump at a chance to replace their entire donné even sooner than I would. That said, just how quickly we would get rid of whatever we now consider our identity is irrelevant. The important question is whether a pour-soi that chooses to change its donné wholesale thereby ceases to be a pour-soi. The answer seems to be no. Even if I choose to move to Mars, become sexless, wear the face of Simone de Beauvoir, speak only Navajo and delete all memories of my Soviet childhood, I will remain in fundamentally the same condition of being conscious and having to interact with the world by making choices. I might no longer be the same person (always assuming that the concept of person is still in use by then), but I will still be a pour-soi.

What is it about a pour-soi that ensures such endurance? If it is not a fixed feature or a cluster of features, it could be a continuous process that involves whatever features happen to be around. As we have seen, that is exactly the option Sartre goes for: a pour-soi is something that happens to the en-soi on which it ontologically depends. In other words, Sartre's idea of le pour-soi is somewhat analogous to the modern view of combustion: while
the ancients took fire to be a substance in its own right, we see combustion as a chemical process involving whatever happens to be burning. Sartre offers two ways of thinking about the process that constitutes *le pour-soi*. On the one hand, he describes what *le pour-soi* does to *l’en-soi* as *néantisation*, i.e. roughly the creation of a gap in the plenum of being, mentioned earlier (e.g. p. 585). On the other hand, one can think of *le pour-soi* as a series of choices. Sartre is quite clear on this point: to be human, he states, “is to *choose yourself*” (p. 587, emphasis in the original). The “only way we ever grasp” our own slippery nature is by thinking about it “as a choice in the process of making itself” (p. 635).  

For the purpose of this investigation, the identification of being human with inescapable choice is the most important feature of *L’être et le néant*. Before I can adopt and put to use a Sartrean view of choice, however, I have to discuss Sartre’s insistence that human choice is a manifestation of what he calls freedom. To be more precise, Sartre’s usage suggests to me that *choix* and *choisir* are simply a way for him to talk about freedom as a process rather than a state, which the noun *liberté* implies. In any case, he equates *réalité-humaine* with *liberté* just as unambiguously as he equates it with choice. Freedom, Sartre says early on, “is not a faculty of the human soul”; it “precedes human nature and makes it possible”; in fact, freedom effectively is human nature since freedom and being human are “impossible to distinguish” (p. 68). Sartre makes this point multiple times. If one thing is abundantly clear by the end of *L’être et le néant*, it is this: “my freedom” as a *pour-soi* “is not an added quality or a *property* of my nature; it is the very fabric of my being” (p. 584, emphasis in the original). As a *pour-soi*, I do not exist first and have the property of freedom later. My existence is the process of making free choices.

The idea that humans are fundamentally free in their choices is perhaps the most controversial part of *L’être et le néant*. As one commentator notes, “there is hardly another example in the history of philosophy” of such a radical insistence on human freedom (Kampits 1995, p. 160, my translation). For Sartre, there is no compatibilism, no middle road between absolute determinism and absolute freedom. A human being is either “entièremen déterminé” or “entièremen libre” (p. 589), and Sartre opts for the latter. This view might seem bizarre, coming as it does from an atheist thinker with Marxist sympathies

---

12 For Sartre, *néantisation* and choice are two sides of the same coin, and both are crucial for his dialectic. For my purpose, however, one side of the coin is enough. As I have hinted earlier, it was Sartre’s focus on choice that drew me to his account of the human condition in the first place. This is why *néantisation* plays no role in what follows.
and very few signs of a Heidegger-style contempt for science. Surely, one might say, we must admit all the myriad ways in which humans are not free. I did not freely choose my birthplace, my first language, my social class or my genome, to name just a few obvious things forced upon me by reality, and it is definitely not my free choice to decay and die, as I inevitably will. True, things might change under fully automated luxury communism or some other utopia, but even if that brave new world does arrive one day, billions of people will still have lived their lives in a vale of unchosen misery, toil and boredom.

If one is to make sense of Sartre’s view of freedom, one should probably begin by conceding that his use of the words liberté and libre is idiosyncratic, not to say unfortunate. To be sure, Sartre clarifies that he uses liberté in its “technical and philosophical” sense of “autonomy of choice” (p. 641). Contrary to “common sense” usage, Sartre says, “to be free” should not be understood as “to get what one wants”. What it means instead is “to make up one’s mind [se déterminer] to want (in the broad sense of choose) by oneself” (pp. 640-641). The distinction between “freedom of choice” and “freedom to get”, Sartre stresses, is “essential” (p. 641). It is not difficult to see, however, how this distinction, illuminating though it is, might not really stop people from taking Sartre’s liberté to mean what it does not. The problem is that any talk of freedom, however artfully defined, lands us in the midst of the free will debate. Once we are there, we are no longer talking about what it is like to be human, or what a human being is in the first place. Instead, we typically take some form of human ontology for granted and then see if humans have a particular property, namely freedom. As we have seen, Sartre tries to pre-empt this by declaring that freedom is not a property but “the very fabric” of being human. Within Sartre’s dialectic, this move works fine, as long as we keep in mind that being human is a process, not a thing, and as long as we then identify this process with making inescapable choices. However, Sartre himself repeatedly muddies the waters by talking about freedom as if it were a property of human choice. Even his terminological clarification discussed above relies on “freedom of choice” (liberté du choix) and “autonomy of choice” (autonomie du choix) (p. 641). Yet the very moment we allow freedom as a property of anything at all, we have to defend it in the same old framework of the free will debate. What exactly makes human choice free? When is human choice not free? Is this freedom genuine or just an epiphenomenon riding on deterministic or random mechanisms that actually drive human choices? Is the idea of freedom even coherent? The free will quagmire that Sartre hopes to escape by equating being human with freedom sucks us back in with a vengeance.
Thus, at least for the purpose of this investigation, Sartre’s account of the human condition is much easier to defend and build on if we dispense with freedom altogether. Consider this statement, variously echoed throughout much of *L'être et le néant*: “Everywhere, the human being [*réalité-humaine*] encounters resistance and obstacles that it did not create; but this resistance and these obstacles only make sense in light of and through the free choice that is the human being” (p. 648). For the most part, Sartre’s observation here is easy to follow. For something to be an obstacle, it has to hinder an agent in achieving a goal. Goals, however, are not a given part of the world; they are set by the agent. What Sartre calls the *coefficient d'adversité* of the things you encounter (see e.g. pp. 638-39) depends on what you want to do with those things. Suppose you were born in Russia and are a Russian citizen. Your Russian passport is a given – it is part of your *donné* – whether you want it or not. Yet it only becomes an obstacle in light of the goals you choose. If your goal is to go to Paris and visit a café where Sartre and de Beauvoir used to hang out, your Russian passport is an obstacle: you need a visa to enter the Schengen Area. If, on the other hand, your goals do not involve you leaving Russia, you may never experience your citizenship as an obstacle. A *pour-soi* has no way of knowing how much adversity there is in a state of affairs until it “freely sets the goal with regard to which the state of affairs is either threatening or favourable” (p. 646).

The one element of Sartre’s observation above that is hard to recognise or defend is the description of human choice in general and human goal-setting in particular as “free”. Translating Sartre’s point into familiar terms of more recent Anglophone debates might help us to see that more clearly. When Sartre talks about choosing your own goals or projects, which then imbue the world around you with meaning, the distinction he relies on is similar to Frankfurt’s first-order and second-order volitions (Frankfurt 1971) or Watson’s desiring versus valuing (Watson 1975). To keep things simple, let us call it a distinction between lower-level desires and higher-lever desires, with no evaluative connotations attached to “lower” or “higher” respectively.

In the Russian passport scenario, your wish to get a Schengen visa is a lower-level desire, while your wish to pay homage to Sartre and de Beauvoir is the higher-level desire that prompts you to want a Schengen visa in the first place. On a Frankfurt-style account, this distinction serves to tell free willed action from unfree willed action: your action is
genuinely free only if you do what you want to want to do. If you succumb to a lower-level desire that goes against a higher-level one – say, if you spend your next holiday drinking in St. Petersburg instead of going to the trouble of getting a visa – then your action is not truly free. A potential problem for such an account of free willed action is infinite regress: it is not clear why the buck stops one level higher rather than two, three, or even more levels higher. To stop the infinite regress, we need to postulate some kind of non-random connection between the agent and her higher-level desires, which is a challenging task (see Andreev 2017 for a detailed discussion).

However, what is a weakness in a Frankfurt-style hierarchy of desires is a virtue for Sartre. Recall that Sartre’s definition of being free is making up your mind “to want (in the broad sense of choose)” by yourself. The possibility of an infinite recursive regress (make up your mind to make up your mind to want to want etc.) is not a threat to Sartre’s definition. On the contrary, an infinite regress of choices is precisely what his idea of freedom seems to depend on. It is safe to assume that Sartre is aware of the obvious objection that – as Schopenhauer’s oft-quoted dictum has it – we are not free to want what we want. For Sartre, however, this objection is belied by the fact that we can ask ourselves whether we do in fact want to want what we want. It is belied even more by the fact that we can keep posing that question ad infinitum. For Sartre, a pour-soi’s inability to ever stop such questioning – its inability to “find any aid, any foothold” for its choices in what has been chosen before (p. 635) – is the very guarantee of its freedom. A human being is ultimately free because its “reality is purely interrogative” (p. 811, emphasis in the original). For every moment of human existence, there is a choice to make, and for every choice, there is a follow-up question about more choices.

Again, as long as we bracket Sartre’s talk of freedom, this observation seems instantly familiar. Allow me an extended illustration of just how pervasive the experience of self-interrogation can be once it has begun. Whenever I reflect on my next action, I get the unshakeable impression that every moment is a moment of questioning the choices I can’t help making. Take this investigation. The number of choices I have to make in order to finish it can be overwhelming. On a bad day, nearly every sentence feels like a tough choice not to give up and do something else. How do I structure this sentence? Which synonym do I use? Do I keep playing this academic Glasperlenspiel? Do I write another column about the moral bankruptcy of the Putin regime instead? Do I want to stick with philosophy? Do I...
want to go back to working on my next novel? Do I want to go back to working on my next novel?

This unrelenting self-questioning is by no means limited to momentous choices. It reaches all the way to the most mundane. There is a cat on the windowsill right next to me. She looks like she wants to be scratched behind her ears. Do I scratch her? Do I continue typing uninterrupted? Every second I don’t scratch the cat – every second I realise I’m not scratching the cat – I can’t shake the feeling that I am making a choice in not scratching her. Perhaps even more strikingly, I can’t escape self-interrogation with regard to my beliefs. Once I realise that I have a belief – say, that there is a cat on the windowsill – I face the choice between continuing to hold that belief and convincing myself otherwise. After all, believing in the existence of a cat on a windowsill is no trivial matter. I might be hallucinating, or I might be misapplying the term “cat”. The cat might be an alien in disguise. Other people have held similar beliefs; why couldn’t I? Even my most fundamental convictions are not immune to questioning and choice. For example, I believe that there is no reason to believe in an immortal reified self. Should I stick to that belief? Or is it just an old habit that I should subject to critical examination?

Let me stress that these examples are not meant to illustrate a feature of human consciousness as such. As I argued in 2.3, consciousness need not be reflective – either in the highly recursive language-mediated way shown above or in any more basic way. However, once consciousness does become reflective and acquires the ability to pose the question “What do I do next?”, there is no way to stop this kind of self-interrogation except by repeatedly choosing not to engage in it. Once I can pose the question “What do I do next?”, the experience of permanent choice accompanied by self-questioning does not need a philosophical defence. Like the experience of plurality of objects or the experience of being here rather than over there, it is an inescapable part of me, the Writer.

The notion of free choice, by contrast, is confusing and superfluous. To be sure, I experience some of my choices as more constrained and others as less constrained. However, I do not experience them as free, at least not in any sense of “free” that could withstand even a handful of textbook arguments against human freedom. In particular, I do not – cannot – experience my choices as “free” in a sense that applies in equal measure to future actions and to past actions. This is because past actions lie outside the participant perspective. When
I look at the world as a participant, my past is just as given, i.e. *donné*, as my body or location in spacetime. As Sartre puts it, the past is “out of reach” and “haunts us from a distance” (p. 656). What is outside my participant perspective is outside my experience.

One can object that the same argument can be levelled against the experience of choice. Do we not talk about past choices as routinely as we talk about future ones? We obviously do. But if so, choice also appears to be outside the participant perspective. Thus, by the same logic, I cannot experience it. My response is to agree with the objection: we cannot experience past choices. Ideally, we should be discussing choice in a language that has a distinct term for choices made in the past and another term for choices that still have to be made. If I were aware of such a language, I would borrow both terms for the purpose of this investigation. As things stand, I have no choice but to introduce an unwieldy noun phrase for the kind of choice that Sartre equates with being human. I will call it *unavoidable future choice*, or UFC. Whenever I speak of choice in the sections that follow, it is UFC that I have in mind. Consequently, my interpretation of Sartre’s account is built specifically on UFC rather than choice in general.

Before I finally sum up my takeaway from *L'être et le néant*, though, one last comment on freedom is in order. I have just introduced a distinction between two kinds of choice: past choice and (unavoidable) future choice. The purpose of this distinction is to show how I can have immediate experience of choice, i.e. how choice can be a feature of the participant perspective. Given my argument above, one may wonder why I do not go ahead and perform a similar conceptual split on freedom, instead of emphatically distancing myself from Sartre’s insistence on *liberté*. After all, the past-future asymmetry that I ascribe to the notion of choice is unmistakably revisionist. As noted above, it is not part of common usage in any of the languages I am familiar with. In for a eurocent, in for a euro, one might say. Since I have entered the business of concept revision anyway, why not extend the same past-future asymmetry to freedom? In other words, why not declare that the experience of being free does not need defending as long as we stipulate that the property of freedom only apply to future actions?

My response is simple: the analogy is a false one. To see that, let us compare a plausible common-sense definition of making a choice with plausible common-sense definitions of being free. For a plausible common-sense take on freedom, we need to distinguish positive
freedom and negative freedom. To be free in a positive sense is to do things that I want and therefore choose to do. To be free in a negative sense is not to do things that something other than me makes me do. Thus, positive freedom is predicated on choice. Even if we allow that there are two kinds of positive freedom – one for past choices and one for future choices – freedom remains superfluous because all it does is track choice. Negative freedom, for its part, relies on the notion of something other than me. In everyday parlance, this is largely unproblematic because the relevant referent of me is typically some part of my person life (see 2.3), specified by the context. However, in a Sartre-inspired account of what it is like to be a self this definition of freedom is a non-starter because most of my person life (my body, my past, my social relations) is part of le donné and therefore something other than me. Consequently, if we were to try to square this definition with Sartre’s view of le pour-soi, then very few, if any, of my choices would ever be free since my choices are clearly a function of my person life.

What about choice though? Why does it make more sense to ascribe to it a past-future asymmetry in light of Sartre’s account? On the most common-sense-friendly definition I can think of, to make a choice is to do one of the things I believe I can do in a given situation. To begin with, this definition is highly Sartre-compatible. It equates choice with action (cf. “Notre description ... ne distingue pas entre le choisir et le faire”, p. 641) and foregrounds the situational nature of choice (cf. “mon choix – le choix que je suis – [est] appréhension de cette situation-là”, p. 726, emphasis in the original). What is more, a Sartre-inspired account does not need to modify this definition by requiring that I be able to do otherwise than I actually do in any strong counterfactual sense. Both in common usage and in a description of the Writer from the participant perspective, I can say that I am making a choice as long as I believe that I have at least two options at my disposal. Demanding any stronger counterfactual ability to do otherwise takes us very far both from common usage and from the actual phenomenology of my choices – much further than the trivial observation that my past choices now feel very different from the upcoming ones. Given all that, I conclude that the common-sense notion of choice lends itself readily to a past-future asymmetry in light of Sartre’s account. The common-sense notion of freedom, on the other hand, resists such revision.

Let me now sum up my takeaway from L’être et le néant as regards describing the Writer, i.e. me, from the participant perspective. First off, here is a brief glossary of my
interpretations of Sartre’s key terms. *L’en-soi* is whatever exists and is not self-aware consciousness. *Le pour-soi* is a locus of self-aware consciousness; in particular, it is *my* consciousness. *La réalité-humaine* is a complex four-dimensional cluster of events, similar to Schechtman’s person life. The *réalité-humaine* that an observer would describe as *me* includes all structures and functions of my organism, including my self-aware consciousness. My *donné* is any part of the world that I interact with, including any part of *me* that is not my self-aware consciousness, such as my body or my past.

Now comes the point that is crucial for my definition of the self. What is it like to be a locus of self-aware consciousness, i.e. a *pour-soi*? As we have seen, it is like constantly making unavoidable choices; it is like choosing a possible course of action in the midst of the part of the world that I have to interact with. Since choosing a possible course of action is tantamount to doing, being me “reduces to doing”; being me “is action” (Sartre 1994, p. 632). Being me is choosing a “goal that reveals the world” (p. 633), i.e. imbues the world with meaning, and then making further choices in light of that goal – whether it is to try to achieve it or to replace it with a different goal.

As I try to zoom in on the minimal, indispensable referent of the first-person singular pronoun as I use it to refer to myself, I can strip that referent of my particular location in spacetime, my particular social connections and my particular past. In fact, I can chisel away any contingent part of the world from that referent by imagining that I live and have lived somewhere else, at another time and with a very different body, speaking other languages. The only thing I cannot eliminate through this Cartesian thought exercise – no matter how hard I try – is my awareness of being something conscious that has to make choices and is indistinguishable from the choices it makes. In other words, the minimal enduring referent of my *I* – the referent of my *I* that I cannot choose not to refer to at least as part of a larger referent once I realise that I *can* choose to refer to it – is a choice-maker acting upon contingent phenomenal content. This choice-maker shapes itself from whatever content is available, without ever becoming that content. One of the few features I can describe as permanently *mine* in me the choice-maker is the fact that my choices are forever open to questioning. Every choice I make leads to more unavoidable choices in the future. Ultimately, I always have to choose without “any aid, any foothold”, i.e. without ever
finding anything beyond myself that can make my choices and fill the world with meaning for me.  

In the previous section, I described the Writer from the observer perspective as a complex event cluster brought about by biological and cultural evolution. I argued that, in order to have a self, such a cluster needs to be distinct, localized, unified, conscious, active, endowed with language, and capable of reflecting on its own properties. In this section, I have tried to show that from the participant perspective the Writer appears to be a series of unavoidable choices, a choice-maker aware of constantly having to choose what to do and what to be next. In the remainder of chapter 2, I draw on a particular hypothesis regarding the function of human consciousness in order to reconcile the two descriptions.

2.5 Bridging the perspectival gap

The last section looked at the Writer from the participant perspective. The participant perspective leads a conscious agent to see some possible future events as courses of action the agent can take. When I, the Writer, adopt the participant perspective and try to identify the minimal enduring referent of each token of the singular first-person pronoun in my self-referring speech, I arrive at a picture of myself that is similar to Sartre’s pour-soi. As a participant, I can best describe the minimal referent of my I as a series of unavoidable future choices. These choices are forced upon me by the totality of the contingencies of my existence. In other words, from the participant perspective I am a choice-maker aware of constantly having to choose what to do next and thus what to be next, in the Sartrean sense of being that never fully coincides with itself. The choices I that I make – the choices I that am – never appear to be final. They are always open to questioning and further choices.

In what follows, I refer to the minimal enduring referent of the Writer’s I described above as a Choice-Maker. The goal of the next section (2.6) will be to reconcile my experience of being a Choice-Maker with the observer perspective in general and an evolutionary

---

13 Cf. Nagel’s rhetorical astonishment in *The View from Nowhere* at having to identify himself with a set of contingent properties: “How can I, who am thinking about the entire, centreless universe, be anything so specific as *this*: this measly, gratuitous creature existing in a tiny morsel of spacetime, with a definite and by no means universal mental and physical organization? How can I be anything so small and concrete and specific?” (Nagel 1986, p. 61, emphasis in the original) What Nagel describes as “the view from nowhere” strikes me as the experience of unrelenting choice taken to its logical conclusion: wherever I choose to be, I can always conceive of choosing to be somewhere else given the opportunity, which means that I can conceive of being everywhere and nowhere in particular.
approach to human consciousness in particular. That will be my final step towards defining the self. If I succeed in identifying a plausible counterpart of the Choice-Maker on the observer perspective, that counterpart can then serve as the definition of the self for my attempt to understand self-knowledge.

The goal of this section is to clarify three points regarding the conceptual gap between the two perspectives and my attempt to bridge it. The first point is why I expect the Choice-Maker to have a counterpart on the observer perspective in the first place. The second point is why translating the concept of choice to the language of the observer perspective appears to pose a particular challenge. The third point is why we can hope to overcome that challenge.

Why should the Choice-Maker have a counterpart on the observer perspective? As a first step towards an answer, recall that in 2.4 I defined the difference between the observer perspective and the participant perspective as the difference between treating something as an event or as a possible course of action. Therefore, the difference is a pragmatic one: it assumes that an agent can describe the same (four-dimensional) part of the world in two distinct ways and then make different use of those distinct descriptions. Thus, I can describe my upcoming holiday in terms of statistical probabilities and their causes (observer perspective); or I can describe it as something I have to make up my mind about one way or the other (participant perspective). The former may help me to understand my past behaviour in similar circumstances as well as teach me something about human behaviour in general; the latter will allow me to make relevant choices and act upon them. Crucially, however, the two descriptions do not concern themselves with two ontologically distinct phenomena. Both target the same part of the world, namely whatever I will end up doing during my upcoming holiday. Summing up, we can say that the relation between the two perspectives is that of two different types of models, developed for different practical purposes but describing the same reality. Therefore, if any part of the world has a description on one perspective, we can assume two things: (1) it can have a description on the other perspective; and (2) there is a way to recognize that the two descriptions deal with the same part of the world.

I have now clarified why I expect the Choice-Maker to have a counterpart on the observer perspective. The second point I need to discuss is why finding this counterpart is not a trivial
matter. After all, in the holiday example above the two descriptions of my holiday plans have such an obvious affinity that it seems self-evident that they describe the same part of the world. Why should describing the Choice-Maker from the observer perspective be any more challenging?

Let us call perspectival translation any act of finding a counterpart description on the other perspective for any (four-dimensional) part of the world. Roughly, the general form of such translation from participant speak to observer speak will proceed as follows: a part of the world that is described by an agent as “I can choose to do action x” can be rendered into the observer perspective as “Event x will\(^{14}\) happen given set of causes y”. Because of the highly complex causality of human behaviour, in practice identifying the relevant set of causes may be difficult or impossible. This, however, does not make most perspectival translation conceptually challenging: as long as we have developed relevant explanatory models, we can apply them to any events, including those that involve human agents. A conceptual hurdle first arises when we try to translate the notion of choice as such. The approach I have just outlined succeeds so easily precisely because it brackets the “I can choose to do” part. However, if our goal is to find a counterpart description for the Choice-Maker, this move will not work. Clearly, it is the bracketed part itself that we need to render into explanatory models formulated from the observer perspective.

This does seem difficult. The reason for the apparent difficulty is that choice, as defined in 2.4 above, is a notion that only makes sense on the participant perspective: to make a choice is to do one of the things I believe I can do, i.e. to treat some future events as courses of action. As long as we accept this definition, the perspectival element of choice is irreducible: there is no choice if there is no locus of consciousness which involves considering alternative futures and deciding on a course of action. Accommodating the idea of alternative futures within the observer perspective is tricky in its own right: it is not clear how future can possibly come with alternatives as long as we subscribe to a naturalist worldview. To observers with naturalist commitments, the world is ontologically similar to Sartre’s en-soi: it is what it is, and it will be what it will be (assuming that tense even applies to its fundamental description). Everything that happens in the world is an event that unfolds

\(^{14}\) We can choose a different modal expression if we don’t want to commit to a deterministic view of causality: x is likely to happen, x might happen etc. Which modal we choose is not important for the general picture of perspectival translation that I sketch here.
in – to use a distinction from Sturma (2015), inspired by Wilfrid Sellars, – the space of causes (*Raum der Ursachen*); while the space of reasons (*Raum der Gründe*), with the corresponding talk of actions and choices, is just a useful way of describing certain events in certain contexts. It does not matter whether we see the world as fully deterministic or partly random. As long as something is a part of the world at all, it happens just like everything else that happens. Consequently, it has no multiple futures\(^{15}\); it only has the future that it happens to have as it continues to happen.

Our scientific edifice is built on the assumption that this ontological uniformity encompasses humans just as it encompasses bacteria, hurricanes or galaxy clusters. While many of the models we use to explain human and animal behaviour do involve the concept of choice, those models are not usually concerned with reconciling the perspectival nature of choice and the observer perspective. As I indicated in my sketch of perspectival translation above, our models simply take it for granted that certain parts of the world behave *as if* they had alternative futures and *as if* they could choose among those futures. This is fully in line with non-reductive, or integrative, naturalism, which assumes that naturalist explanatory models do not have to be mutually translatable in order to be successful. In other words, even if the experience of having alternative futures and of choosing among those futures (which, I have argued, is inescapable and does not need a dialectic defence) is largely or fully epiphenomenal for a molecular biologist or a physicist, this does not mean that a psychologist or a political scientist have any reason to worry. After all, we do not need a physical or biological account of the causal role of anger or democratic elections in order to make meaningful use of those concepts in relevant contexts. Likewise, we do not need to attribute any biological or physical causality to choice when we talk of choice in psychology or political science.

Having said that, the domain of this investigation is human ontology. Ontology looks at what exists at the highest – or, if we prefer a different metaphor, the deepest – level of abstraction. It trades in the most fundamental. Should we worry about translating choice to observer speak when we build ontological models? Recall that I have identified the minimal enduring referent of my *I*, viewed from the participant perspective, as a Choice-Maker.

---

\(^{15}\) I understand that this holds even on the Many Worlds interpretation of quantum mechanics. Regardless of how many alternative realities are associated with a wave function, the wave function itself remains what it is. See e.g. Carroll 2019.
Recall also that choice is irreducibly perspectival; it is not clear how we can possibly work it into a physical, chemical, biological or even neurological model. Yet if we cannot embed choice in any such model, then we cannot ascribe to choice any causal efficacy in the world described by fundamental science. But if so – if there is no way to talk of choice and causal efficacy in the same breath – it seems that, as far as our most fundamental scientific models are concerned, the experience that constitutes Choice-Makers has to be an epiphenomenon. In other words, at the level of description that is fundamental enough to bear on ontological considerations our choices as choices do not seem to cause any events whatsoever. No matter how we translate them to observer speak, we end up placing ourselves in the space of causes, where we have no role to play.

I have now discussed why finding a counterpart of the Choice-Maker on the observer perspective may seem like a daunting task. We have seen that we may indeed worry about treating choice merely as a model useful in certain contexts, and that one likely reason for our worry is this: we want our choices – the choices that we are, as Sartre would put it, – to have genuine causality all the way down or at least much of the way down, cutting across different theoretical models and levels of description. With that concern articulated, I can now address the third and final point that requires clarification: why, despite this apparent hurdle, we can hope to succeed in rendering the Choice-Maker into observer speak. In a nutshell, my answer is to propose and then dismiss a likely reason why it might seem important for our choices as choices to cause things regardless of the model or level of description.

The likely reason in question is suggested by L'être et le néant. In L'être et le néant, as I tried to show in 2.4, Sartre consistently conflates the concept of choice with the concept of free choice and, by extension, with the idea of freedom as such. Sartre’s concern, recall, is not simply to show that we constantly have to choose what to do next. Rather, one of his goals is to convince us that, in some profound ontological sense, we are free because we constantly have to choose what to do next. This conflation of unavoidable future choice and freedom, I argued, is hard to defend conceptually. Perhaps even more importantly, it is not warranted by phenomenological data: while I do experience my choices as more or less constrained, I do not experience them as free in any profound ontological sense. However – and this is crucial for the question at hand – the conflation of choice and freedom can help to explain why we might want our choices to have causality as choices even on the observer
perspective. If we consistently think of choice as *free* choice, then our worry about its causal powers is ultimately pegged to our worries about freedom. We might feel that, as long as our *free* choices are not genuinely causal across (much of) the board, we cannot take any ultimate credit for our achievements; nor can we ultimately blame and punish people for doing bad things.

If that is the case – if we want our choices to have genuine fundamental causality on the observer perspective because we want to have genuine fundamental freedom, robust enough to justify both our pride and our vindictiveness, – then my investigation can offer no remedy. I do not know how to defend *free* choice, and I am not even sure that the idea of such robust freedom is coherent. If, on the other hand, we care about the causality of choice to the extent that we care about the causality of its counterpart – whatever it may be – in observer speak, then translating choice may be no more or less worrisome that any other act of translation between different models that describe the same part of the world but cannot be reduced to each other. Recall that, by its very definition, unavoidable future choice (UFC) has nothing to do with pride or vindictiveness, i.e. with credit or responsibility for what has already happened. The domain of UFC is what is about to happen. Thus, if we are careful enough to stay focused on UFC and to keep it free of freedom, we can hope to bridge the perspectival gap and find a plausible counterpart of the Choice-Maker on the observer perspective. This is the goal of the next section.

2.6 The self as a reflective matrix of maybe

In 2.3, where I first considered the Writer from the observer perspective, I argued that the self was best understood as a function, and I stipulated a number of conditions that any candidate self-function must fulfil. I described the Writer as a distinct, unified, localized experiencer-agent that interacts with other such agents and has a capacity for self-reference, which in turn presupposes self-awareness and a sign system such as language. Consequently, I assumed that the self-function must be a function of such experiencer-agents. If the Writer has a self that lends itself to a definition on the observer perspective, this self must serve some purpose, since serving a purpose is what a function is.
What purpose could that be? The goal of this section is to identify a specific empirically attested function of biological organisms that meets the conditions stipulated in 2.3. The phenomenological evidence discussed in 2.4 will provide a crucial clue in this effort.

2.4 (“The Writer from the participant perspective”) was an attempt to show that from the participant perspective the minimal enduring referent of I is a series of unavoidable future choices, which I called a Choice-Maker. The weight of this phenomenological evidence compels me to assume that, whatever the self-function may be on the observer perspective, it must – at least in part – be structurally similar to making choices. An additional clue is one of the conditions stipulated in 2.3, namely the assumption that selfhood, unlike consciousness, is limited to biological organisms that self-refer using a sign system such as language. Whatever function the self performs, that function must be easy to identify in linguistic animals that have reached a certain level of maturity, and it must be harder to identify in other animals. Summing up both clues, we have reason to expect the self-function to be something above and beyond the mechanisms of maintaining organismic homeostasis that are found in all conscious organisms, including very young human children and non-human animals. These mechanisms include a number of features involving consciousness: proprioception, a unified body schema, a sense of body-world boundary, a spatiotemporal centredness of perception, and so on. As I argued in 2.3, such features are necessary for a self to emerge, but they do not presuppose a self; nor do they amount to a self.

Let us start with the phenomenological clue. The self-function, as stipulated above, must have structural similarity to making choices. Recall the definition of making a choice from 2.4: to make a choice is to do one of the things I believe I can do in a given situation. In particular, focus on the “things I believe I can do” part of this definition. My belief that there are things I can do presupposes a belief in multiple possible futures. As a choice-maker, I necessarily see the future as holding alternative options. This fact bears on my choices in two ways. On the one hand, as I argued in 2.5, it is a necessary condition for choice in the first place. On the other hand, imagining alternative possible futures plays an important role in the self-questioning that accompanies my choices because it often helps me to justify preferring one course of action to another.

To be clear, we are not simply talking about reacting to contingencies on the basis of available information. Note again the “I believe” part of the definition of choice offered
above: to make a choice is to do one of the things I believe I can do in a given situation. The notion of choice presupposes an awareness of mutually exclusive possible courses of action before the choice is made. For some choices, this awareness may be very brief; for others, it may last for days or even years. For some choices, it may come and go as a single vague mental image; for others, it may be verbalised and rich in detail. Whatever the exact phenomenology of such awareness, it must have some phenomenology. This is because the experience of making a choice has a particular temporal structure. First, I become aware that there are at least two mutually exclusive things I can do in the immediate or more distant future. Then, I do one of those things (note that doing here includes mental actions such as making up your mind). By contrast, if I simply react to contingencies without ever holding in mind alternative futures, I do not seem to engage in choice. Instead, my behaviour is hardly distinguishable from a knee jerking in response to tapping.

Thus, the phenomenological clue leads me to conclude that the self-function must have something to do with imagining mutually exclusive futures. Let us now turn to the other clue, which is based on the link between selfhood and linguistic self-reference. The self-function, the second clue suggests, should be easy to find in linguistic animals, such as human children that have reached a certain point in their development. By the same token, it should be harder to find in other animals, including human children that have not yet reached that point. Can we identify a specific function of biological organisms that meets the criteria suggested by both clues? An answer to this question requires a foray beyond the core of the MAP – this time not into phenomenology but rather into the realm of empirical research. In other words, our question is this: do we have empirical evidence that after a certain age human children outperform non-human animals and younger human children in tasks that require imagining mutually exclusive futures?

As it happens, we do. Let me start by relaying one particular piece of evidence; the reason for choosing it will become clear in the next paragraph. Redshaw and Suddendorf (2016) report an experiment involving 90 human children aged 2-4 as well as eight adult great apes (3 chimpanzees and 5 orangutans). In the experiment, the subjects were presented with a forked tube, shaped as an upside-down Y. A small object (a ball for the children or a grape for the apes) would be dropped into the top opening of the tube. The subjects could then try to catch the object by covering one or both of the bottom openings with their hands. Covering both openings at once, needless to say, ensured a 100% success rate. In the human
group, most of the 4-year-olds and many of the 3- and 3.5-year-olds covered both openings on the very first trial and then consistently applied the same strategy across all trials. None of the 2-year-olds covered both openings on the first trial, and many of those 2-year-olds who did eventually cover both openings reverted to covering only one opening on later trials. In the ape group, only one of the subjects ever covered both openings at all; she then reverted to covering one opening at a time. When the main experiment was over, some of the apes were offered extra attempts. This was done to rule out the possibility that the apes lacked the dexterity to cover two openings at once. Two apes eventually learned to cover both openings by trial and error, i.e. occasionally regressing to the less successful strategy. Most of the 2-year-olds who were offered extra attempts eventually learned to cover both openings in the same way.

I mention this particular study because it is cited by Baumeister et al. (2018) among other evidence in support of their view of mature human consciousness as a matrix of maybe. According to Baumeister et al., a key function of mature human consciousness is to imagine “multiple alternative possibilities” and evaluate them “with the help of anticipated emotional outcomes” (2018, p. 235), i.e. in light of other imagined alternative possibilities. The older human children, they argue, did so well in Redshaw and Suddendorf’s experiment because they were quickly able to “understand the future as multiple different maybes” (p. 225). By contrast, the adult apes did not appear to think of the future as a forking path. Even though they were obviously modelling what was about to happen (otherwise they would have had little interest in covering either of the bottom openings), they seemed unable to model two mutually exclusive futures at the same time.

Baumeister et al. frame their view as a theory of prospection, i.e. thinking about the future. To begin with, they assume that the origins of prospection are pragmatic: we think about the future in order “to prepare for how to act” in a way that steers us “toward desired outcomes and away from aversive ones” (2018, p. 224). The ability to infer what is about to happen and then act upon that inference has a clear evolutionary pay-off. This ability, however, is evident in many non-human animals. Human prospection, Baumeister et al. stress, goes well beyond projecting a causal sequence into the immediate future. What is special – and especially powerful – about human prospection is that it is thoroughly nonfactual: it allows us to “see beyond what is there” (2018, p. 224), translating contingencies into sets of possible futures. Human consciousness, Baumeister et al. suggest, “creates alternative
possibilities” (2018, p. 224). As far as I can tell, they mean that both factually (“in the sense that it makes it possible for them to become reality”, p. 224) and conceptually, i.e. in the sense that the very idea of a possibility presupposes a conscious creature thinking of the future within what Baumeister et al. call the matrix of maybe.

Parts of Baumeister et al. (2018) provide such striking parallels to L'être et le néant that one almost expects to find Sartre on the list of references. While Sartre is not there, the authors do make an explicit nod to phenomenology and “the concept of ontological horizon”, i.e. “the set of possibilities … given one’s present situation” (p. 225). “Ontological horizon” sounds esoteric enough to give a reader steeped in modern Anglophone philosophy a whiff of the continental, but it might be helpful to note some more transparent similarities between Baumeister et al.’s matrix of maybe and its phenomenological precursors.

For instance, the idea that possibility “comes into the world through la réalité-humaine” plays a crucial role in L'être et le néant (Sartre 1994, p. 161). So does the idea that being human – whether we identify it with freedom, choice or just being conscious – is fundamentally relational: we only exist “in relation to a state of affairs and in opposition to (malgré) that state of affairs” (Sartre 1994, p. 643). As Baumeister et al. express this idea, all action that we prepare for “occur[s] in the context of a contingency matrix” (2018, p. 225). Our consciousness can only generate a matrix of maybe, i.e. a set of possibilities, because it exists within a set of contingencies, i.e. what Sartre describes “my body”, “my place”, “my past”, and so on. A third parallel between Baumeister et al. and Sartre that is worth mentioning here is the idea of meaning as a function of making choices in light of other choices. For Sartre, humans can only “discover” the meaning of their immediate situation and of their world in general by choosing a (larger) goal (une fin) in light of which they can evaluate their other choices (see e.g. Sartre 1994, pp. 633, 646, 655-656). Likewise, Baumeister et al. appear to equate the meaning of a possible action with its evaluation in light of “anticipated emotional outcomes” (2018, p. 235). Meaning, they say, “inherently locates entities in context of alternative possibilities” (p. 235). No course of action, one might say, makes sense except in a matrix of maybe.

Given the parallels I have just listed, one might also reasonably wonder why I bother to supplement my takeaway from Sartre with Baumeister et al.’s matrix of maybe in the first place. Is it just because Sartre is a French existentialist – and therefore fundamentally
suspect in an analytic context – while Baumeister, Maranges and Sjåstad are modern psychologists, who can lend a veneer of empirical respectability to essentially the same idea? There is, I must admit, an element of that. However, it was certainly not my main motivation. Recall that in 2.4 I drew on Sartre’s account because I needed to describe the Writer from the participant perspective. Identifying the minimal enduring referent of my I with a Choice-Maker, i.e. a series of unavoidable future choices, captures what it feels like to be me, but it does not explain how we are to understand that experience in an evolutionary framework. By contrast, Baumeister et al.’s account points to a way to bridge the perspectival gap, i.e. a way to make sense of the Choice-Maker from the observer perspective. In particular, Baumeister et al.’s view of prospection suggests a perspectival translation of choice which has causal efficacy. In what follows, I take the final step towards a definition of the self from the observer perspective by sketching how a conscious matrix of maybe can be involved in causality at the relevant level of description.

To show how a conscious matrix of maybe can have a causal role, all we need to grant is that consciousness in general is causally efficacious. This assumption does not strike me as far-fetched.\textsuperscript{16} For one thing, conscious brain activity appears to involve a number of brain regions and take a lot of time and energy that could be spent on other functions important for survival. “It is implausible”, Baumeister, Lau et al. argue (2018, p. 2), “to assume that such a costly adaptation would have been selected for” if either the organism or the species did not stand to benefit from it. Secondly, there is always the immediate (meta)evidence of the causal efficacy of consciousness provided by the fact that we have linguistic labels for conscious phenomena and write papers discussing them. Thirdly, there is a way to account at least for some types of causality involving consciousness in standard reductionist terms. While we may not be able to reduce the intrinsic what-it-is-like-ness of experience to neurological and biological concepts, we can reduce “similarities and differences” between conscious phenomena \textit{functionally}, e.g. by showing how they affect “our cognition and capacity to cope with the world” (Kim 2007, pp. 416-417).

Once we have granted that consciousness in general is causally efficacious, we can focus on the causal efficacy of consciously simulated alternative futures. Baumeister et al. clearly

\textsuperscript{16} I cannot hope to begin to do justice to the issue of the causal efficacy of consciousness in this paper. My brief sketch of the reasons for assuming that consciousness is causally efficacious is only meant to show that I do have such reasons and that those reasons are independent of the goal of this investigation.
assume that prospection in general and prospection involving multiple alternative futures in particular cause behaviour. As we have seen, their key idea behind the matrix of maybe is that considering imagined scenarios makes an organism more likely to survive and reproduce. If imagined futures serve a biological function, they should be functionally reducible to neuronal activity. How would such reduction proceed?

Luckily, this is not a hypothetical question. There has been a wealth of work on the mechanisms of mental time travel in general and prospection in particular. The emerging consensus is that “remembering past events and imagining future events depend, to a very large extent, on shared cognitive and neural processes” (Addis et al. 2008, p. 2222). As far as the brain’s functioning is concerned, we simulate episodes of the future the same way we simulate episodes of the past. This makes perfect sense from a pragmatic point of view. Presumably, the survival value of having any kind of memory, however basic, of an event or a type of event is in the fact that it helps the organism to deal with similar events in the future. In pragmatic terms, retrieving a past event is no different from picturing a future one; both serve to influence future action. To put it in the terms of this investigation, both serve to facilitate what the agent experiences as unavoidable future choice. Thus, as long as we grant causal efficacy to memory, we should accept that prospection is causally efficacious in exactly the same way. From the observer perspective, there is no theoretically significant difference between causal sequences involving multiple retrieved memories and causal sequences involving multiple imagined futures.

Here, then, is a perspectival translation of the notion of choice into the observer perspective: what I experience as an instance of choice is an event that involves conscious simulation of alternative future events as part of its causal structure, understood in terms of Mackie’s INUS model (Mackie 1980). From this translation, it follows that what I experience as the minimal enduring referent of my I, i.e. the Choice-Maker acting on the contingencies of my person life, is a series of events that involve conscious simulation of alternative future events as part of their causal structure. To put this in terms of Baumeister et al.’s theory of prospection, what I minimally refer to when I use the singular first-person pronoun to refer to myself is a conscious matrix of maybe functioning within a contingency matrix.

Now that I have equated myself the Choice-Maker with a neuronal mechanism simulating multiple alternative futures, I have all but completed my final step towards a definition of
the self. The only thing that remains to do is to emphasize self-reflection and self-reference as necessary elements of selfhood. Recall that the stipulation of self-reflection is based on the view that self-consciousness is necessarily reflective consciousness, which I adopted in 2.3, as well as on the stipulation of self-reference. The stipulation of self-reference, in its turn, is based on the semantic affinity between the self and I, assumed in 2.2, and on the rules for using the singular first-person pronoun, considered in the same section. With that in mind, here is a definition of the self:

The complex function that we can call the self is a matrix of maybe operating within a contingency matrix and reflecting on its own operation.

It is worth stressing that this definition does not require that the self, thus defined, have in mind any particular “self-theory” when it self-refers or reflects on its own workings. A self-referring matrix of maybe may think of itself as a matrix of maybe, or a human animal, or a brain, or an immortal Christian soul, or Quetzalcoatl the serpent god, or the universal self-luminous Ego of the Upanishads. In other words, a self may lack some or much of the knowledge that it could, in principle, have about itself. What it means for a self to have or lack such knowledge is the subject of the next chapter.
Chapter 3. Self-knowledge

3.1 Knowledge

The goal of my investigation is to understand what makes self-knowledge apparently special. Therefore, my chief concern is what difference, if any, obtains between self-knowledge and other knowledge. Defending, let alone developing, a particular theory of knowledge as such goes beyond the scope of this essay and is above my current pay grade. Having said that, I need to make at least some assumptions about what knowledge is in order to talk about how self-knowledge is special. In this section, I adopt a definition of knowledge and then briefly discuss some of the assumptions it is based on.

Here is the definition:

An organism knows that $p$ if and only if it has appropriately acquired a disposition to act as if $p$ were true, and $p$ is true. Consequently, knowledge that $p$ is an appropriately acquired disposition to act as if $p$ were true, given that $p$ is true.

Here is a quick explication of this definition in terms of what Alvin Goldman called “the post-Gettier knowledge-trade” (quoted in Becker, n. d.). The view of knowledge I adopt for the purpose of this investigation can be described as naturalist, externalist, process-reliabilist, and contextualist. It is a form of epistemic naturalism because it assumes that the conditions for having knowledge involve “natural relations like causation and reliability” (Poston, n. d.) rather than purely logical relations such as implication or coherence. It is a form of epistemic externalism in that it assumes that the conditions for having knowledge are at least in part external to the knower. It is a form of process reliabilism as it requires that whatever natural process creates “the knowledge-constituting link between belief and truth” (Becker, n. d.) be appropriately reliable; i.e. it must “produce a high ratio of truths to falsehoods” (Goldman & Beddor 2016). Finally, it is a form of contextualism since it allows that ascribing knowledge may depend on the context of knowledge ascription; “different contexts” may “set different epistemic standards” (Black, n. d.).

The rest of this section pursues two auxiliary goals. Firstly, it motivates and thereby clarifies the link between knowledge and action, stipulated by the definition above and crucial for
this investigation. Secondly, it attempts to show how one can be a contextualist without embracing a kind of epistemic relativism that might threaten our entire scientific enterprise, analytic philosophy included. In doing so, it highlights a key distinction between reflective and non-reflective action; it also makes a point of distinguishing knowledge from certainty.

To start with knowledge and action, there are at least three reasons to link the two. One of these reasons is internal to this investigation; the other two are independent of this investigation. As regards the internal reason, recall that the goal of this essay is to make sense of how self-knowledge is special in terms of its connection to agency. Consequently, construing knowledge as such through its relation to action rather than in terms of belief can save us some dialectic steps.

The first of the two independent reasons is the evolutionary origin of cognition. As the definition above suggests, I take knowledge to be a relation between a biological organism and the way the world is, where “the world” fully includes the organism itself. If knowing something is a relation involving the world and a biological organism, then the ability to form such relations must have evolved. Presumably, it evolved to facilitate a certain kind of action, namely action aimed at survival and reproduction (even if some species may use the ability to acquire knowledge in ways that are not apparently linked to either). Indeed, if we try to make sense of reliabilism in evolutionary terms, it seems plausible that a reliable way to form “the knowledge-constituting link between belief and truth” (Becker, n. d.) was originally one that led to successful action; and successful action was action that helped the organism to survive and produce offspring. It is, in any case, difficult to think of a different reliability criterion to apply to a time before language and philosophy.

The other independent reason for linking knowledge to action is the way we acquire meta-knowledge. To begin with, it is worth noting that knowledge is not always available for reflection and resulting meta-knowledge. An organism can enjoy a knowledge relation to the way things are without knowing or even being able to know that it has such a relation. However, whenever knowledge does becomes available for reflection and resulting meta-knowledge, it first manifests itself as action.17 For example, in order to know that I know how to ride a donkey I have to ride donkeys. Likewise, in order to know that I know that the

17 Here it is worth stressing again that this essay uses a broad definition of action: anything that an organism does on purpose, including whatever it mentalizes on purpose, qualifies as action.
African wild ass is the wild ancestor of donkeys I have at least to have thoughts to that effect.

Those were my reason for defining knowledge through its relation to action. Now I turn to contextualism. My goal here is not to defend a contextualist view of knowledge per se. Instead, I aim to justify my adoption of a form of contextualism in the context of writing a philosophy paper. In the rest of this section, I consider an apparent threat that epistemic contextualism poses to the very practice of (analytic) philosophy, and I offer a way to diffuse that threat which will have implications for my understanding of self-knowledge.18

Admittedly, it might not matter much for the purpose of this investigation what specific view of knowledge I adopt. As long as it is a view attested and defended in the literature, it should be able to do the job I need it to do, i.e. provide a basis for contrasting knowledge as such with self-knowledge. Epistemic contextualism, however, is somewhat of a special case. As shown by Coliva (2012, pp. 54-59), adopting contextualism in a philosophical context seems to have unwelcome implications precisely because the context is that of philosophy.

Generally, a contextualist approach to knowledge assumes that the conditions for ascribing knowledge may vary depending on the context of knowledge ascription (Hübner 2015). For example, they might depend on what is at stake in potential action informed by the knowledge in question. Suppose that you tell me that on Saturdays my local Systembolaget outlet is now open until 3 p.m. rather than 2 p.m. You believe that because you looked at the opening hours the last time you went there a month ago. On a contextualist view of the kind described, whether you know when Systembolaget closes depends on what is at stake for me if I choose to rely on your words. If I just want to pop in and get a bottle of wine to fill the empty spot in my wine rack for the sake of symmetry, I will be inclined to say that you do know when the shop closes. If, however, I am throwing a big party and have not yet purchased any alcohol beverages at all, I will be less inclined to say that you have knowledge.

18 In what follows, my understanding of contextualism is based on Coliva (2012), Hübner (2015), Ichikawa and Steup (2018), and Van Loon and Sarzano (2016). Having said that, any misunderstanding and confusion are solely mine.
In philosophical contexts, Coliva (2012) notes, such a take on knowledge leads to an apparent paradox. To show how her reservations bear on my discussion, I will put them in the terms of the definition of knowledge above. In philosophy, to act as if $p$ is, among other things, to assert that $p$ in academic journals, at conferences, in MA theses, and so on. To be able to say “S knows that $p$” (as opposed to “S argues that $p$” or “S holds that $p$”) in such fora, we typically want S to have acquired her disposition to act as if $p$ in a way that is reliable enough for any context, and we tend to apply the same stringent criteria of knowledge ascription to our own beliefs. As a result, the moment we embrace contextualism as philosophers we suddenly appear to both know and not know a whole lot of things. On the one hand, in our everyday lives we have plenty of appropriate reasons to act as if $p$ with regard to thousands of propositions, ranging from “It is snowing outside” to “Proxima Centauri is the closest star to the Sun”. Thus, in everyday contexts it appears that we can ascribe to ourselves and others knowledge of all those things. In philosophical settings, however, we may be very reluctant to state that we know any of the very same propositions, especially when we start debating scepticism and relativism. Somehow, most of our knowledge is downgraded to mere belief by our own philosophical yearning for unassailable context-independent justification.

How then can we do philosophy, embrace contextualism and save our knowledge from deteriorating into mere belief all at the same time? French philosopher Jacques-Henri Vollet suggests one way out of this quandary (Vollet 2016). When it comes to knowledge and action, Vollet observes, we have three plausible but apparently incompatible intuitions:

1. Knowledge is not sensitive to the practical situation of the subject.
2. There is a normative relation between knowledge and action.
3. The epistemic conditions necessary for acting on a proposition vary with the practical situation of the subject. (Vollet 2016, p. 5, my translation)

Suppose that we want to keep all three intuitions. How do we do that without either relativizing knowledge or denying the link between knowledge and pragmatic concerns? Vollet offers a two-pronged solution. On the one hand, he distinguishes two kinds of action: reflective and non-reflective (réfléchie et non réfléchie). In non-reflective action, the agent “responds to the reasons [for the action] directly or spontaneously”; a dog barking at the postwoman is one example. By contrast, in reflective action the agent’s “response is
mediated by a conscious and explicit evaluation of the reasons”. The “paradigmatic example” Vollet offers is “weighing the pros and cons before acting” (Vollet 2016, p. 9).19

The other part of Vollet’s solution is to distinguish knowledge and certainty. In cases of non-reflective action, it is rational for the agent to act upon $p$ if and only if she knows that $p$. When it comes to reflective action, however, it is rational for the agent to act upon $p$ if and only if she is certain that $p$. If we accept this dualist epistemic/doxastic norm for action, Vollet argues, we can do justice to all three intuitions listed above by contextualizing not knowledge ascription as such but rather the distinction between knowledge ascription and certainty ascription.

As is normal for any substantive proposal in philosophy, Vollet’s view comes with inherent difficulties (see Van Loon & Sarzano 2016 for a detailed discussion). For instance, in many pragmatic situations it may not be clear whether the action should count as reflective or non-reflective. In other situations, the degree of reflection involved in the action appears to track how much is at stake for the agent, which would seem to make the distinction between reflective and non-reflective action superfluous. Having said that, I find Vollet’s proposal both plausible and helpful as stated. Firstly, the very idea of (un)certainty seems to presuppose reflection, so linking the two in a doxastic norm for reflective action is highly intuitive. Secondly and crucially for my purpose, Vollet’s account helps to solve the apparent paradox described by Coliva (2012).

The paradox, recall, is this: if we make knowledge depend on pragmatic considerations, then as philosophers we end up simultaneously knowing and not knowing most of what informs our actions in non-philosophical contexts. Now, if anything at all can be described as reflective action, it is surely philosophical action such as asserting that $p$ in academic journals, at conferences, in MA theses, or indeed in your own head. To put this in terms of prototype semantics, in the head of a Western philosopher philosophical action must be to reflective action what table is to furniture. By Vollet’s lights, this means that philosophical action requires a high degree of epistemic certainty to be rational. In fact, one can argue that the degree of epistemic certainty required to assert unequivocally that $p$ in a philosophy journal or in one’s own philosophy-plagued head is largely unattainable because any reasons

19 To put this in terms of recent Anglophone epistemology, non-reflective action is triggered by propositional justification, while reflective action requires doxastic justification.
to believe that \( p \) are vulnerable to further sceptical reflection. None of this, however, means that I do not know many or even most of the things that inform my day-to-day actions, or that it is irrational for me to act based on that knowledge.

To make this point clearer, here is again my earlier definition of knowledge, followed by a definition of certainty phrased in similar terms:

**Knowledge:** An organism knows that \( p \) if and only if it has appropriately acquired a disposition to act as if \( p \) were true, and \( p \) is true.

**Certainty:** An organism is certain that \( p \) if and only if it has considered reasons for and against acting as if \( p \), found reasons to act as if \( p \) and found no reasons to act as if not \( p \). (As elsewhere in the essay, “to act” here includes “to engage in thought”.)

In the second definition, note the absence of any reference to an appropriate kind of reasons or to the way things are. Unlike knowledge, certainty easily lends itself to an internalist definition. Certainty is not a relation between the organism and the way the world is, including the subset of the world that constitutes the organism’s reasons for acting as if \( p \). Instead, it is a relation between the organism and only that particular subset. To be sure, certainty may co-occur with knowledge and vice versa. Indeed, given that we cannot ascribe knowledge from any other vantage point except our own, we can only hope that in the long run there is a positive correlation between the two. However, neither presupposes the other. Knowledge does not presuppose certainty because it can be non-reflective (and certainty is necessarily reflective); and certainty does not presuppose knowledge because organisms – including philosophers – can be dead certain of \( p \) even when \( p \) is not part of the way things are.

How does that help us with Coliva’s paradox? For one thing, if we disentangle knowledge and certainty from each other it seems relatively easy to stomach the idea that, as philosophers, we are both certain and uncertain of many things, depending on what we want to do with those things.\(^{20}\) Secondly, the discussion above seems to suggest that reflection is

---

\(^{20}\) For instance, in the context of writing a newspaper column or having a kitchen conversation I can ascribe to myself complete certainty that any form of theism is wrong or that (political) libertarianism fails as a strategy for ensuring a wide range of personal choice for as many citizens as possible. In the context of a philosophy
not inherently rational. While reflection may positively correlate with rationality, my reading of Vollet’s account implies that too much reflection may harm rationality because it demands an irrationally high level of certainty to perform certain actions, such as to ascribe knowledge of the way things are to a cognizing organism in a philosophy journal. In other words, it leads to what Marconi (2007) calls “the overdramatization of truth” (la drammatizzazione della verità), a result of too many thinkers conflating certainty with knowledge and philosophical questions with questions in general.

To conclude, I have argued that contextualism, implied by the link between knowledge and action, is defensible in a philosophical context. I have defended a form of contextualism regarding the relation between knowledge and action by relying on two distinctions, suggested by Vollet (2016): that between reflective and non-reflective action and that between knowledge and certainty. I can state my takeaway from Vollet for my definition of knowledge as follows. Defining knowledge as having a particular kind of disposition for action need not make knowledge unstable or relative. What varies from context to context is the level of reflection and corresponding certainty required for a particular action. The more we reflect on the reasons for a given action, the more certain we may need to be before we act. It is to be hoped that reflection-based certainty correlates positively with knowledge, but it does not imply knowledge. This is because certainty is an internal relation between the organism and its reasons for action, while knowledge is an external relation between the organism and the way the world is, including the organism’s dispositions to act in a certain way and their causal history. All of this has implications for the next section, where I finally try to make sense of self-knowledge in light of my account of the self.

3.2 What it means to know yourself

In the previous section, I adopted the following definition of knowledge:

An organism knows that \( p \) if and only if it has appropriately acquired a disposition to act as if \( p \) were true, and \( p \) is true.

As for the self, at the end of chapter 2 I defined it as follows:

---

seminar at Uppsala University, however, I would not claim to be certain about either of those things, and I would subject anyone claiming such certainty to counterarguments.
The complex function that we can call *the self* is a matrix of maybe operating within a contingency matrix and reflecting on its own operation.

In other words, I identified the minimal enduring referent of (every token of) *I* in my self-referring speech with a reflective mechanism for simulating, evaluating and choosing among alternative futures. As its input, this mechanism uses whatever evidence is available to the human animal I happen to be. Call this view the Self as a Reflective Matrix of Maybe, or SRMM. Given the definition of knowledge above, to know that SRMM is true is to have appropriately acquired a disposition to act as if SRMM were true in a world where SRMM is true. Now, suppose that SRMM is true in the actual world. Suppose also that our disposition to act as if it were true has been acquired in an appropriate fashion. What we need to clarify then is this: what does it mean to act as if SRMM were true?

Let us begin by recalling that 3.1 introduced a broad distinction between reflective action and non-reflective action. In non-reflective action, the agent acts on a set of reasons without reflecting on alternative courses of action. A typical example would be someone accidentally touching the ironing plate of a hot iron and immediately withdrawing their hand to avoid getting burned. By contrast, in reflective action the agent compares alternative courses of action before going ahead with one of them. A typical example would be someone buying a new iron after spending an hour agonizing over the mind-boggling selection of irons in a modern capitalist retail outlet.

Section 3.1 also followed Vollet (2016) in distinguishing two norms for rational action, depending on the type of action involved: an epistemic norm and a doxastic norm. For non-reflective action based on *p* to be rational, the agent must know that *p* is true. For reflective action based on *p* to be rational, the agent must be certain that *p* is true. The degree of certainty required for reflective action may vary with the context, and it may increase with the amount of reflection involved. For example, in order to write the sentence “I therefore conclude that the view of the self as a reflective matrix of maybe is correct” in a philosophy paper and simultaneously describe that action as rational in light of our common search for truth rather than her academic career or other personal goals, a modern Western academic philosopher might need a degree of certainty that is unattainable within the philosopher’s lifetime.
However, making claims in philosophy papers typically constitutes only a tiny fraction of all action that a human animal has to engage in. Many non-reflective or marginally reflective actions that we perform on a day-to-day basis would likely be hampered if we treated them as genuinely reflective. As I first noted at the end of 3.1, this suggests that from a practical point of view reflection as such is not inherently rational. Let us define as rational any action that helps to achieve a goal set by the agent. Thus, if I am an agent and my goal is to iron my shirt before joining a Zoom meeting ten minutes from now, any action that brings about that goal is rational in light of that goal. Consequently, it might be more rational for me to treat the requisite chain of actions (set up ironing board, get iron, plug in iron, put shirt on board etc.) as non-reflective. Indeed, reflecting on things such as the exact placement of the ironing board in my room or whether I am using the most efficient ironing technique would be decidedly counterproductive and therefore irrational.

Given SRMM, what does all this mean for action involved in self-knowledge? To answer this question, it might be helpful to make a provisional distinction between theoretical self-knowledge and practical self-knowledge. Let us define having theoretical self-knowledge as having a true theory of the self. Trivially, in a world where SRMM is true, theoretical self-knowledge manifests itself through actions such as thinking thoughts and writing papers to the effect that SRMM is, or is likely to be, true. In such a world, expounding a different theory of the self (“The self is a pure impersonal Ego”, “The self is a soul”, “The self is an illusion”, “The self is a human animal” etc.) is perfectly possible, but it does not constitute an act of self-knowledge.

It is when we turn to practical self-knowledge that the implications of living in an SRMM-world arguably become more interesting. Let us define practical self-knowledge as any linguistic act of truthfully ascribing a linguistic predicate to a self-referring sign, such as “I” or “Kostia Andreev”, or any non-linguistic act that the agent can explain linguistically by truthfully ascribing a linguistic predicate to a self-referring sign21. For example, if it is true of me that I know English and that I like Icelandic indie music, then my saying things like “I

21 It is worth stressing again that by predication I always mean linguistic predication, as in “‘Like Icelandic music’ is the predicate in the sentence ‘I like Icelandic music’”. I assume that linguistic utterances containing predication can be true or false.
speak English” or “I love Icelandic indie music” is an act of practical self-knowledge. By the same token, if I select the English subtitles when watching an Israeli show on Netflix or if I buy a ticket for the Iceland Airwaves festival, those are also acts of practical self-knowledge. I can explain them, respectively, as “I chose the English subtitles because I know English” and “I like Icelandic indie music; therefore I’m going to Iceland Airwaves”.

Suppose that I have embraced SRMM and now regard the self as a series of unavoidable choices among alternative imagined futures, made in contingent circumstances. Suppose also that I want to perform an act of practical self-knowledge. To stick with the same examples, I may want to ascribe to myself apparently true predicates such as “know English” or “like Icelandic indie music”. As noted in the previous paragraph, making a true claim about yourself is an act of practical self-knowledge. By the very nature of making claims, this involves reflection, i.e. consideration of alternatives. Earlier we saw that such reflective action requires doxastic justification in order to be rational. In other words, in order to make a claim such as “I speak English” or “I like Icelandic indie music” I ought to be certain that those claims are true. However, it follows from SRMM that such certainty is not only unattainable; it is, in fact, conceptually impossible. This is because, on the view of the self as a reflective matrix of maybe, any evidence I may consider when making a claim about myself is not evidence of me. Instead, it is evidence of the contingent circumstances I have to act upon. For example, I may have copious amounts of evidence apparently bearing on the claims “I speak English” or “I like Icelandic indie music”. Yet all this evidence pertains to some contingent part of the world that I have to interact with. It is, in other words, evidence regarding the human animal that generates me; or else it is evidence regarding the person life that has unfolded so far as a result of the interplay between me and my contingency matrix. However I choose to describe it, it is not me. Instead, it is something I have to take a stance on by making yet another unavoidable choice.

By this point at least, one should start worrying whether SRMM, adopted in conjunction with the discussion of knowledge above, might end up undermining itself. Let us go back to the distinction between theoretical self-knowledge and practical self-knowledge, which I made earlier in order to dismiss some seemingly trivial implications of SRMM. It should

---

22 Here and in the rest of this section, I omit for the sake of convenience the specification that the action in question is a result of an appropriately acquired disposition. However, this specification should always be borne in mind, along with the assumption of an SRMM-world.
now be clear that it is not a distinction between two co-hyponyms. Rather, theoretical self-knowledge is a special case of practical self-knowledge. A human agent cannot act as if claims such as “The self is a reflective matrix of maybe” or “The self is a series of Sartrean choices” only applied to something other than herself. Her participant perspective forces her to interpret them as first-person statements and therefore acts of practical self-knowledge: “I am a reflective matrix of maybe”, “I am a series of Sartrean choices”, etc. However, like any other first-person ascription, such a claim would be reflective and thus require certainty to be justified. At the same time, the sheer degree of reflection involved in choosing such a claim over possible alternatives (“I am a pure impersonal Ego”, “I am a soul”, “I am an illusion” etc.) suggests that the requisite degree of certainty may be unattainable.

In other words, SRMM itself appears to preclude certainty about SRMM. As soon as I identify myself with a series of unavoidable future choices, I have to take a stance on that identification. When I ask myself if I should keep choosing to believe SRMM, reasons to reject it readily present themselves. To name just one of them, I can choose to convince myself that SRMM is just a naturalist-sounding reissue of all those pure transcendent entities from the Upanishads and Plato’s dialogues; that it is, in other words, yet another exercise in wishful thinking, motivated by the perennial human desire to detach oneself from this “unclean” contingent world and thereby from the pain and absurdity of the human condition. After all, SRMM alienates in one fell swoop every single contingent property I may be said to have except the very process of simulating, evaluating and choosing among contingent alternatives. As a result, with SRMM I get both to transcend this messy reality and to avoid downgrading myself to a mere illusion, while maintaining a semblance of doing analytic philosophy with respectable naturalist commitments.

It might seem then that my investigation has come a full circle and found itself back on square one. Rather than explain the apparent agency-related specialness of self-knowledge, SRMM risks exposing itself as just another manifestation of that specialness. Moreover, it is conceivable that the very evidence I marshalled for the agency-related specialness of self-knowledge might point to psychological rather than philosophical issues. If so – if the proneness to thought experiments alienating oneself from all contingent features of one’s person life is best explained as a deviation from a psychological norm – perhaps a philosophical enquiry is not the best way to address the underlying issues. One may wonder whether a therapist’s couch or more physical exercise could do the job better.
There is, however, at least one promising way to salvage SRMM as an account of the agency-related specialness of self-knowledge. The first element of this rescue strategy requires us to take SRMM seriously. To take SRMM seriously, we need to go back to Sartre’s no-own-features account of consciousness. On Sartre’s account, recall, no feature of reality that I the *pour-soi* am aware of is a feature of me. Instead, it is a result of my consciousness “reaching out to”, i.e. interacting with, some feature of the world – such as a location in spacetime, a body, or a memory – without ever coinciding with that feature. Accepting Sartre’s view of consciousness allows us to bite the bullet and accept that hardly any evidence bearing on first-person ascriptions allows us to ascribe any features to the self. Consequently, no knowledge I can acquire on the basis of such evidence is evidence for predicates describing *me*. Instead, it is evidence for predicates describing the world, including the human animal that generates me and the person life that unfolds as a result of my acting upon the world. Thus, the vast majority of actions involved in what I have called practical self-knowledge are not instances of self-knowledge at all. Claims such as “I speak English” or “I like Icelandic indie music” are not claims about me; they are claims about the part of the world that forms my contingency matrix. The only self-knowledge in the proper sense is what I termed theoretical self-knowledge: a disposition to think SRMM-believing thoughts and expound SRMM in various contexts, using language with the appropriate amount of hedging.

To sum up the previous paragraph, taking SRMM seriously means that no first-person ascriptions except for first-person statements of SRMM itself are instances of self-knowledge. Rather, they are manifestations of knowledge about the contingency matrix within which the self operates. As a second step on the way to salvaging SRMM, we need to recall that reflection is not inherently rational, and we need to consider the implications of that observation.

As defined earlier in this section, to be rational an action must help to achieve a goal set by the agent. Reflection is a form of action. Specifically, it is a form of action involved in what we can call reflective action, i.e. action in which the agent considers mutually exclusive courses of action before choosing one of them. Clearly, the best way to achieve a goal is to choose the best available way of achieving it. This is why, as we have seen, reflective action requires certainty to be rational: you need to be certain that you are choosing the best course
of action in light of your goal. However, the extent to which it is rational for an action to be reflective in the first place – the extent to which it is rational to consider alternatives before acting – may vary from context to context. In some contexts, a high degree of reflection may be too time-consuming. In other contexts, it may even be useless – for instance, when the agent lacks the evidence or any non-random criteria for choosing among the available options.

These considerations, coupled with SRMM, finally suggest a way to make sense of the apparent agency-related specialness of self-knowledge. Here is how the reasoning can go. A human animal that has reached a certain stage in its development is equipped to reflect on its own functioning. In other words, it can engage in self-reflection. Thanks to the ability to simulate, evaluate and choose among alternative futures – i.e. thanks to nonfactual pragmatic prospection, described by Baumeister et al. (2018) as the matrix of maybe, – self-reflection may go on indefinitely. This is because the matrix of maybe makes any self-ascription vulnerable to a realization that “things could be otherwise”. Assuming that the overarching goal of the average human is to get on with their life, such a high level of reflection is often irrational. In many cases, we do not have any non-contingent criteria – “any aid, any foothold”, as Sartre might put it – for preferring one self-ascription to another.

There are at least two common ways of addressing this problem. One is not to engage in self-reflection in the first place. Another is to become certain of the non-contingent nature of some set of self-ascriptions and act accordingly. For many humans, however, neither of those strategies seems attractive or indeed possible. Luckily, there is an alternative. As a human animal, I can try to identify whatever non-contingent features of me still remain once every single contingent feature has been peeled off by self-reflection. This is how I arrive at SRMM: a view of the self as a matrix of maybe functioning within a contingency matrix and capable of reflecting on its own operation. If I then take SRMM seriously, I realize that most of what we usually describe as self-knowledge is not knowledge of me; it is knowledge of my contingency matrix. Any value I assign to \( x \) in \( I(x) \) immediately threatens to detach itself from \( I \) because the minimal enduring referent of \( I \) is not part of the contingency matrix. Instead, it is the very mechanism that turns the rest of the world – including the human animal that generates that mechanism – into a contingency matrix.
To sum up, self-knowledge appears to be special because most of our true first-person claims do not have the form I PREDICATE. Whatever predicate we are talking about, the general form of such claims is, roughly, *I have to act in a world where C PREDICATE*, where C is the relevant part of the contingency matrix. As for knowing that C PREDICATE, there is nothing special about that knowledge in terms of its relation to our agency. The *I have to act in a world where* part stays the same regardless of what C PREDICATE refers to.
Conclusion

In this essay, I have tried to understand why, no matter what I say about myself, it feels like I may be talking about something other than myself. The essay has argued that one possible reason is just that: most of the time when we think we are talking about ourselves we are really talking about something else. The only thing we can say about ourselves is that each of us is a series of inevitable choices. I have identified that series of choices with a reflective matrix of maybe: a self-aware mechanism of pragmatic prospection working with whatever it has to work with.

In other words, this essay has argued that each of us is a chain of the same steps repeated over and over and over again. Step 1: take stock of the world. Step 2: think what you can do with that world. Step 3: choose one of the things you think you can do. Repeat for as long as you exist because you can’t choose not to make choices. There is no other self-knowledge. Anything else we say when we think we are talking about ourselves is part of the first step. Anything else we think we know about ourselves is knowledge of some part of the world. It may be something about the human animal which keeps you – you the reflective matrix of maybe – in existence. It may be something that has happened in the life of that human animal. Whatever it is, it is not you. Instead, it is something you have to deal with.

Adopting this view of self-knowledge does not necessarily make it easier to deal with life choices. It can, however, be helpful or indeed liberating in at least three distinct ways.

To begin with, it might be easier to let go of certain first-person ascriptions if we think of them as features of the world rather than of ourselves. For example, if we are privileged to live in a society that allows a relatively high degree of personal choice in matters like place of residence, occupation, personal relations, lifestyle, language, gender, physical appearance and so on, we do not have to cling on to any of those features simply because we have clung on to them before.

At the same time, embracing such a view of self-knowledge might also help us to abandon the belief that some features of the world are a matter of personal choice. We often assume that human agents can change certain features of their physical or psychological make-up
simply because we have always ascribed those features to the self. However, every self makes choices within a contingency matrix, i.e. within the particular set of circumstances the self has to deal with. Any feature of the human animal that keeps the self going is part of that set of circumstances in the same way as any other feature of the world. Just like any other feature of the world, any feature of a human being may be easy to change, or it may be difficult to change, or it may be impossible to change within the human being’s lifetime. Consequently, we might need to know which is which rather than simply expect ourselves and others to make certain choices.

Finally, the view of self-knowledge defended in this essay might help us to see more clearly that in many cases it does not matter which choice we make; or else we have no way of knowing whether it matters. If every single moment of our lives is a moment of choice – if each of us is made up of a myriad of inevitable choices, some potentially momentous but most extremely trivial, – then there can hardly be anything special about any given choice as such. We have to choose which choices are important enough to agonize over and which can be made routinely and non-reflectively.

I choose to hope to explore these and other implications of the view of the self as a reflective matrix of maybe in the future.
References


de Waal, F. (2016). *Are we smart enough to know how smart animals are?* W. W. Norton & Company.


https://doi.org/10.1007/978-3-476-05427-2

https://oll.libertyfund.org/titles/hume-a-treatise-of-human-nature


http://www.gutenberg.org/files/57628/57628-h/57628-h.htm#Page_317


https://www.oxfordscholarship.com/view/10.1093/0198246420.001.0001/acprof-9780198246428


https://doi.org/10.1016/j.cub.2016.04.062


https://doi.org/10.1057/9780230513389


https://doi.org/10.2307/2024703