This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Embodied Metacognition:

How We Feel Our Hearts To Know Our Minds

John Dorsch

PhD in Philosophy
The University of Edinburgh
2022
Declaration of Authorship

I, John Dorsch, declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

John Dorsch, 12th Jun 2022
Abstract

The aim of the present work is to make a plausible case for the phylogenetic origin of self-knowledge, one which is compatible with a prevalent view about its ontogenetic origin, the social-scaffolding view. Essentially, the phylogenetic origin is generally argued to be evaluative metacognition, i.e. a system of cognitive control mechanisms, while the ontogenetic origin is generally argued to be mindreading, i.e. cognitive capacities supporting mental state attribution. So put simply, the present work aims to provide a plausible solution to the problem of how to establish metacognition as a significant precursor to mindreading.

Previous attempts to establish metacognition as an evolutionary precursor to mindreading have struggled to account for the exact relevance that metacognition has for the origin of self-knowledge. This present work analyzes this struggle as first having to solve ‘the problem of triviality’, explain how metacognitive mechanisms are more sophisticated than mechanisms enabling basic cognitive feats. Once this problem is solved, however, the problem develops into having to show that the added layers of sophistication warrant construing metacognition as a significant precursor to self-knowledge. This latter problem is analyzed as the ultimate problem, referred to here as simply ‘the problem of the origin of self-knowledge’. Thus, the present work seeks to offer a plausible solution to both of these problems.

The solution to these problems is argued to be embodied metacognition, cognitive control mechanisms that enable conscious self-control of information-seeking behavior through the production of strongly embodied feedback signals. These feedback signals, often called ‘noetic feelings’, are theorized to motivate, demotivate, and guide cognitive acts to be more accurate. Here it is argued that these signals are strongly embodied, characterizing conscious emotional states constituted by psychological imperatives, states with external manifestations in the form of bodily expressions that can be detected, monitored, and regulated by individuals in the subject’s socio-cultural environment.

To argue for this, I augment evaluative metacognition in light of evidence from recent interoception-based research in metacognition which reveals the degree to which noetic feelings are strongly embodied. They are thus argued to be cerebrally generated feedback signals produced by cognitive comparators that become intricately intertwined with bodily processes. Recent
evidence suggests these signals come to fulfill their role in cognition through the conscious perception (the interoception) of internal bodily changes, specifically cardiovascular changes, and hence the subtitle of the present work, “How We Feel Our Hearts to Know Our Minds”.

Reconceptualizing evaluative metacognition as embodied metacognition provides us with the tools for solving the two problems above. Concerning the first, mechanisms enabling conscious self-control of cognition are clearly more sophisticated than mechanisms that enable basic feats of cognition, such as reinforcement learning. Moreover, appealing to strongly embodied noetic feelings offers the possibility of solving the first problem in such a manner that the second problem is also solved. As conscious expressions of emotions, noetic feelings demand robust information-seeking behaviors and manifest feelings of confidence and uncertainty externally, thus they are well-suited to link metacognition to the social-scaffolding view about the origin of self-knowledge.
Famously, the phrase “know thyself” was once etched above the entrance to the Temple at Delphi. Though it was likely meant as an admonishment to the worshippers entering there, a reminder of their status as frail and flawed mortal subjects, the Greek maxim was, and still is today, often given a wholly different meaning. Whether famously or infamously, Socrates popularized the maxim to mean that each one of us has a duty—to members of our community and to ourselves—to examine the choices made in life by adopting a critical stance towards our own beliefs. The essence of Socratic philosophy is the question: ought we to believe the things we believe?

Indeed, this method of critical thinking might well lay the foundation for any form of reasoning, whether it be philosophic, natural scientific, or otherwise. This is in variably because when it is well applied, this method arrives at justification for belief. So strongly was Socrates committed to the maxim that he was convinced that the very possibility of living a good life and thriving as a human being was predicated upon actively seeking out justification for belief, assuredly justification for others, but more importantly, justification for ourselves, for our own personal wellbeing, so that we come to know ourselves. “The unexamined life is not worth living”, Socrates proclaims upon being sentenced to death for practicing his method of inquiry. But how did we come to acquire the ability to examine our lives in this manner?

This would have been impossible were it not for the fact that we humans possess some capacity to know ourselves, or, at the very least, come to acquire beliefs about ourselves. But how did we come to acquire self-beliefs? The present work starts with a simple assumption: the question about the origin of self-knowledge admits of a natural-scientific answer, one which is comprehensible to creatures like us. As such, we ought to be able to turn to the empirical sciences of the mind and arrive at conclusions about how humans acquired the capacity for self-knowledge.

In developing a plausible story about the origin of self-knowledge, I endeavor to build a stable bridge that extends backward in evolutionary time, one that links the sophisticated cognitive capacities of humans in their preverbal stage of development to the less sophisticated capacities of non-human animals. The bridge I thus tread along is constructed out of the raw materials found in the brain sciences, in particular it is composed through the fusing together of distinct disciplines, gathering evidence from comparative and developmental psychology, based on theories and
models from cybernetics and cognitive neuroscience. But before I introduce the bridge, let’s briefly discuss these key materials.

**Developmental Psychology**

Researchers in developmental psychology believe to have discovered an ontogenetic origin of self-knowledge. Studies on the cognitive capacities of children, from early infancy to school-age, lay the foundation for a highly influential view about how the capacity to ‘know thyself’ emerges. We are not born into this world as self-knowing; rather, we become self-knowing agents gradually over the course of our development and enculturation. First introduced by the Soviet psychologist Lev Vygotsky, this view, sometimes called the ‘social scaffolding view’, will be treated as a starting point for the present work, one of the abutments upon which I shall build my bridge.

Accordingly, self-knowledge is argued to be a product of capacities for ascribing belief states and then inferring implications of those states, a process we learn to do. Evidence suggests we first learn to ascribe belief states to others, without that is, really knowing what we are doing, and then, due to in part the behavior of others ascribing belief states to us, we learn to ascribe states to ourselves, at which point, we are in a position to acquire knowledge about ourselves.

Thus, this cognitive capacity, dubbed ‘mindreading’, is believed to emerge sometime between ages 3 and 5, when children began to pass well-known false-belief tasks and the famous mirror test. The conclusion of this body of research can thus be summarized as this: self-knowledge emerges as the result of specific socio-cultural practices around treating individuals as epistemic agents, as individuals responsible for what they believe. But how, this work sets itself the task of addressing, did these socio-cultural practices emerge?

**Comparative Psychology**

Meanwhile, by deploying the so-called ‘uncertainty monitoring paradigm’, researchers in comparative psychology believe to have discovered a phylogenetic origin to self-knowledge. By issuing tests of an animal’s ability to cope with and mitigate uncertainty, difficult tasks of perceptual discrimination, comparative psychologists believe to have evidence that certain animals, those
whose performance is on par with that of humans, possess cognitive capacities for a rudimentary form of self-knowledge which, they believe, ought to be construed as evolutionary precursors to our own highly sophisticated capacities. This family of rudimentary mechanisms and capacities is often dubbed ‘metacognition’: the animal’s cognitive system possesses mechanisms for monitoring and controlling its own cognition, a form of regulation enabled by the production of feedback signals that inform about the animal’s own cognitive states.

Is there a connection between mindreading and metacognition? And if so, how exactly do these two fit together? Does metacognition somehow prepare the way for mindreading? Or are they ultimately distinct? What exactly are we seeking to explain whenever self-knowledge is our explanandum? Though researchers in developmental psychology tend to be skeptical of construing metacognition as a precursor to mindreading, comparative psychologists are often convinced this story is on the right track and so have, over the span of three decades, developed more and more sophisticated tests for demonstrating the link. But despite this progress, many researchers continue to be reluctant to accept metacognition as a precursor to mindreading. Thus, more work, both of an empirical and of a conceptual nature, needs to be done before the two stories can be united.

**Uniting the Two Stories**

The present work takes on the various issues that have been raised about metacognition and analyzes them into two substantive problems and two corollary problems, setting it as its goal to provide a plausible solution to each one in light of both the empirical and rational evidence, and thus establish metacognition as a proper precursor to mindreading: the initial problem of triviality, its related problems of metacognitive biodiversity and honeybee metacognition, and the ultimate problem of the origin of self-knowledge. By the end of the dissertation, I will hope to have shown a plausible way forward, a sound approach to solving these problems and making inroads into uniting the two stories, and thus uncovering the natural origin of self-knowledge.

As of now, the story about the origin of self-knowledge for creatures like us, which fuses together evidence from both developmental and comparative psychology, admits of the following sketch, a story that has recently been carefully arranged and conscientiously told by experimental psychologist Steven Fleming in his book “Know Thyself: The New Science of Self-Awareness”. Mechanisms for cognitive control create capacities in individuals for regulating their own cognition.
Taken in aggregate, behaviors around regulating cognition in this individualistic manner transform into practices for regulating the cognition of others within the socio-cultural environment. Evolutionary pressures ensure those individuals who are better at regulating themselves in this manner are more likely to survive, thus socio-cultural practices centered around regulating the cognition of individuals become learned as the practices they ultimately are, namely epistemic practices, aimed at beliefs as beliefs, and thus metacognition sets the stage for mindreading.

One crucial issue with this story lies in the middle: how do behaviors around regulating an individual’s own cognition, as basic as these are, transform into the highly-sophisticated socio-cultural practices aimed at regulating the cognition of other individuals? It is still unclear how exactly these cognitive control mechanisms are so advanced so as to produce these sophisticated behaviors, since largely these mechanisms are little more than those which are deployed to explain basic cognitive feats, like adaptive behavior and reinforcement learning. Thus, I refer to this problem as ‘the problem of triviality’, since, if it is not properly addressed, it renders capacities for metacognition little more than basic cognitive capacities, and thus trivial, while also rendering it incapable of solving the ultimate problem of the origin of self-knowledge.

Moreover, this story is unclear in terms of details around how metacognitive mechanisms produce practices for socio-cultural regulation of an epistemic kind, that is, as practices that deal in beliefs as beliefs. Consider just how sophisticated such behavior is. It is one thing to possess beliefs; under some plausible definition, all animals possess beliefs: arguably, my cat, Socrates, comes in from outside because he believes food has been set out for him. But it is a wholly different thing to deal in beliefs as beliefs, as mental entities that might fail to accurately represent the world. If metacognition is made possible by the same mechanisms that make adaptive behavior possible, then it is unclear how this plays a crucial role in giving rise to forms of thinking or behavior that treat beliefs as beliefs. Thus, the story above, despite how well it captures the two disciplines of psychology, leaves unclear how metacognition ought to be construed as a crucial precursor to robust epistemic practices.

To put that succinctly, if one wishes to reconcile evidence of the ontogenetic origin of self-knowledge with evidence of the phylogenetic origin, while simultaneously doing justice to both disciplines, one will inherit a conceptual debt that must be repaid in two separate installments on pain of philosophical default. First, one will need to explain how these cognitive control mechanisms are more sophisticated than mechanisms enabling adaptive behavior. All animals are capable of adaptive behaviors, but not all animals have robust capacities for exercising control over cognition and behavior. To do so would be to solve the problem of triviality. Second, the added
layers of sophistication need to explain how epistemic practices develop, namely practices around treating individuals as holders of beliefs, which implies explaining how beliefs are handled as such by a socio-cultural community. Having the conceptual liquidity to pay back both of these installments means building the bridge I endeavor to draft the blueprints for, one that has the tools to explain the origin of self-knowledge for creatures like us.

The framework of the bridge, I argue, is embodied metacognition, a family of cognitive control mechanisms, like those described by the cybernetic theories at the heart of the phylogenetic origin of self-knowledge, that has been configured by natural evolution, as well as the animal’s own learning history, to produce cognitive feedback signals that guide the animal’s behavior, making it more adaptive and improving its fitness. Crucially though, the feedback signals that do the real interesting work in metacognition are those that are produced in an extracerebral format, in particular through bodily processes involved in the regulation of the cardio-respiratory system, and hence the subtitle of this dissertation: “How We Feel Our Hearts to Know Our Minds”. I argue that mental states characterized by these signals, often called ‘noetic feelings’, ought to be construed as crucial enablers of the more cognitively sophisticated mental states characteristic of self-knowledge.

To set about arguing for this, the present work discusses recent empirical evidence suggesting that noetic feelings have the potential to be strongly embodied, intricately intertwined with extracerebral processes, becoming enablers of online control of behavior and executive functioning. Strongly embodied noetic feelings thus characterize conscious states that subjects use to guide their cognitive behavior, not only whether to act, but also how to best to act. Based on recent studies on noetic feelings and interoception (i.e. the internal perception of extracerebral changes), I argue that the mechanism responsible for the production of noetic feelings consists of an intricate interface between cerebral and extracerebral processes. Thus, construing metacognition as strongly embodied in this manner is crucial to the solution I propose to the problem of triviality, as this mechanism is highly complex and the behaviors it enables are indeed highly sophisticated forms of self-control.

What is more, if the social-scaffolding view is on the right track, strongly embodied noetic feelings have the potential to solve the problem of triviality in such a manner that the problem of the origin of self-knowledge is also solved. A cognitive system that produces strongly embodied noetic feelings ensures metacognitive feedback signals will become conscious signatures of emotion, and this state of affairs has two crucial consequences for making inroads into solving the problem of the origin of self-knowledge. On the one hand, it means that the mental states characterized by strongly embodied noetic feelings will be motivators of rich information-seeking
behaviors, often called ‘epistemic foraging’ behaviors. This is due to how, as feeling states, noetic feelings tend to demand psychologically satisfying resolutions, in which e.g. negatively-valenced feelings of uncertainty, from the perspective of the subject, need to be replaced by positively-valenced feelings of confidence. This psychological demand implies that strongly embodied noetic feelings are well-suited to promote rich epistemic foraging, in which subjects explore and exploit sources of information, behaviors which could plausibly serve as the foundation for socio-cultural practices that forge a collective epistemic agency.

On the other hand, the production of strongly embodied noetic feelings means that individuals possess the means to express internal confidence levels with respect to their information-seeking behaviors through highly nuanced and externally observable bodily reactions that can be readily detected, monitored, and regulated by other individuals in the socio-cultural environment. For our part, humans can almost immediately recognize certain facial expressions, gestures, and even gaits as signifying confidence, or lack thereof. Thus, the present work hypothesizes that externally observable behaviors corresponding to internal levels of confidence serve as crucial planks in the bridge leading from metacognition to mindreading.

In summation, the present work argues that noetic feelings admit of various ways in which the extracerebral body gets involved in their production, and those which are strongly embodied ought to be construed as both implying highly sophisticated cognitive capacities for exercising self-control, as well as construed as crucial steppingstones on the path to developing capacities for acquiring self-knowledge. This is because mental states characterized by these conscious feelings are produced by neurocomputational activity, inform about the adequacy of cognitive processing, are crucial enablers of executive functions, and, as such, strongly embodied noetic feelings, as well as the mechanism producing them, offer a plausible solution to the problem of triviality.

Furthermore, strongly embodied noetic feelings can serve as crucial ingredients to solving the problem of the origin of self-knowledge if the social-scaffolding view is on the right track about its ontogenetic origin. This is because of how noetic feelings promote robust behaviors around acquiring information and are expressed through externally observable bodily reactions that can be easily targeted by practices aiming to regulate behavior. In particular, strongly embodied noetic feelings of confidence characterize the states that ought to be construed as precursors to self-knowledge states. Since confidence, when well placed, approximates competence, any community that learns to regulate behaviors around bodily expressions of confidence will not only incur a massive survival advantage, but will also be engaging in proto epistemic practices not unlike those demanded by the Greek maxim, “know thyself”.

ix
# Contents

## Introduction


2. Are the Noetic Feelings Embodied? The Case for Embodied Metamnmonic Cognition: p.65


4. The Case for Embodied Metacognition: Shoring up Metacognition against the Problem of Triviality: p.118

5. Evaluative Metacognition and the Predictive Processing Framework: Grappling with the Origin of Self-knowledge: p.142


## Conclusions

References: p.196
Introduction

How did we come to acquire the capacity to introspect and reflect, to reason and feel confident? How did the capacity to ruminate, regret and feel ashamed emerge on this planet? In other words: how did it come to be that we humans know that we know, or, at the very least, believe we do? The present work starts with a simple assumption: the question about the origin of self-knowledge admits of a natural-scientific answer, one which is comprehensible to creatures like us. As such, we ought to be able to turn to the empirical sciences of the mind and arrive at conclusions about how humans acquired the capacity for self-knowledge, in particular to developmental and comparative psychology. But before delving into this, let us first spend some time discussing the form of self-knowledge that serves as our ultimate explanandum.

Start by contrasting two scenarios: (1) feeling an itch versus (2) feeling confident. In the scenario (1), you feel an itch, and, let’s say, you promptly scratch it. Here there is an extremely weak form of self-knowledge, one that does not require any capacity to reason and reflect. You feel an itch. You scratch. One might construe this behavior as resulting from various beliefs about the self, e.g. the belief that you are itchy, the belief that the itch is at some location on the body, etc. Though this view might be a hyper-intellectualization of the scenario (I’m inclined to think it is), such beliefs do not properly describe the form of self-knowledge I am interested in explaining, as it is deeply unclear how such beliefs are products of reasoning and reflection.

Consider now scenario (2): feeling confident, in particular confident that you have provided the correct answer to a question. Let’s say, it is a quotidian question about the location of nearest post office. “Where might I find the post office?” “Oh, you’re in luck. There is one just across the street here.” “Really? Are you sure? I don’t see one.” “Yes, I’m sure. See? It’s right there. See the red building? That’s the post office.” Step back and wonder for a moment at the kind of self-knowledge that is beating away here at the heart of this everyday scenario.

Upon responding you are sure of the location of the nearest post office, you have an understanding that there are facts that settle the matter as to whether your belief about the location is true. By saying you are sure, you are ensuring your interlocutor thereby that your belief, and the
utterance that expresses it, is supported by these facts, that your utterance and the belief it expresses are true. By saying you are sure, you express that you believe your belief about the location of the post office is true. Essentially, you have a *metabelief*, a belief about your belief.

There is a technical notion in cognitive psychology for the capacity to entertain metabeliefs such as these, namely ‘self-directed mindreading’, a strange-sounding name for sure, but one that describes entertaining beliefs about your own beliefs as on par with reading your own mind, or at least attempting to do so (see Wimmer and Perner 1983; Perner 1991; Perner and Lang 1999; Perner and Dienes 2003). Some researchers believe it is possible for us to fail to read our own minds (e.g. Carruthers 2011). For example, after a very stressful day at work, we might find ourselves lashing out at a loved one, thinking falsely that they are the cause of our stress. In such an event, we turned our mindreading capacities upon ourselves, attempted to read our own mind, acquired the belief that our loved one was to blame, but we were dead wrong in this belief, and so failed to read our own mind. We are frail and flawed creatures after all.

As humans, we are avid mindreaders, of ourselves, and of others. We commonly ascribe beliefs and other mental states to ourselves, other people, to animals, even to inanimate objects: I, for one, have a strong belief that my computer hates me. But this capacity develops quite late in development. In our culture, in which beliefs are common commodities, mindreading seems to emerge around the age of 4 (see Wellman and Bartsch 1988; Flavell et al. 1990; Zaitchik 1990), just around the time children start to pass the famous mirror test, a common test of self-awareness (Bertenthal and Fischer 1978; Anderson 1984).\(^1\) What is more, tests of mindreading present often insurmountable challenges for even the cleverest of non-human animals (see Tomasello et al. 2003). All of this suggests that self-directed mindreading is made possible by highly sophisticated capacities for obtaining highly robust forms of self-knowledge, precisely the kind of self-knowledge, about which I shall endeavor to tell a natural origin story.

1. Terminological Clarifications

This is a good point to flag some key terminology that will be with us throughout the entire dissertation. Mindreading, as just introduced, is also routinely called ‘metacognition’. This is due to

\(^1\) In the mindreading literature, there is emerging evidence that children younger than 4 are capable of passing modified false belief tasks. Children as young as 15 months are able to pass non-verbal, violation-of-expectation based false belief tasks (Onishi and Baillargeon 2005).
how mindreading encompasses capacities for cognizing about cognition or thinking about thinking (Carruthers 2009a, 2009b). When you entertain a metabelief, you entertain a belief about a belief, a thought about a thought. Due to how mindreading requires the representation of beliefs as beliefs, as mental entities that can fail to describe the world accurately, mindreading has also been called ‘metarepresentational metacognition’ (Perner 2012; Carruthers 2014). This is because entertaining a metabelief requires capacities for representing mental representations as such, that is, as mental entities that can fail to accurately, or appropriately, represent the actual state of affairs about the world (see Proust 2010).

Consider again the belief about the location of the post office. This belief will turn out to misrepresent the actual state of affairs so long as, obviously, the post office is not at the location expressed by the belief. Entertaining the metabelief that the belief about the location of the post office is true (or false) requires capacities for entertaining beliefs as beliefs, as mental entities that can misrepresent the actual state of affairs about the world. Thus, mindreading requires capacities for representing representations as representations, commonly referred to as ‘metarepresentational metacognition’. Going forward, that is what I shall mean by deploying the term ‘metarepresentation’ except when explicitly stated otherwise (such as in Chapter 3).

This manner of representing has a technical notion associated with it. In the analytical tradition in the philosophy of mind, it is called the ‘de dicto mode of reference’, characterized by representing something as the thing it is (Chisholm 1976; Keshet and Schwarz 2019). This fits with the above scenario, since here the belief is represented as a belief. This manner of representation ought to be contrasted with another, distinct, and less robust, manner in which something can be represented, the so called ‘de re mode of reference’ also called ‘indexing’ (Recanati 1993). In the de re mode, something is representing in an extremely basic sense, in which it is merely pointed to, or indexed.

For example, consider the index in the back of any textbook. Say, the word ‘insight’ is indexed to pages 4, 12, 23, etc., and so, in a sense, these indexes represent, or stand in for, the particular passages that feature the word. But these indexes do not represent the passages as passages; rather, the passages are represented as page numbers. Not that there is something crucial about representing something with numbers. Rather, representing passages as page numbers is an

---

2 The term ‘mental representation’ is one of these notions in philosophy that has an extremely long history and has thus been used to mean various things in different contexts (see Pitt 2020 for a review). Unless otherwise stated, I shall use the term ‘mental representation’ or ‘representation’ to refer to those mental entities, like beliefs, that have the potential to be accessible for conscious report and which represent the world in some way. If I deploy the notion to refer to features of brain states that are determined by their computational role (see Marr 1982), I will instead refer to these as ‘subpersonal mental representations’ or ‘subpersonal representations’ unless the context already makes this clear.
example of indexical reference, in which passages are represented, not as what they are (i.e. as passages), but though indexes that stand in for them, pointing, as it were, to them. This distinction between *de dicto* and *de re* reference will turn out to be extremely useful in making sense of what is required at the cognitive level of description for the possession of capacities for self-knowledge.

Since our explanatory target, our explanandum, is the natural origin of distinctly-human self-knowledge, the main character of our origin story should by a kind of holy trinity of mindreading, metarepresentational metacognition, and *de dicto* mode of mental representation. Each of these ultimately picks out the same set of capacities that can be used to acquire self-knowledge, which, in the aggregate, might as well be called ‘introspection’. But I shall continue to use these technical notions, since each highlights a specific gloss, an aspect, or a nuance of introspection, and I shall commonly refer to all as ‘mindreading’.

That said, there is an equally influential conception of metacognition that is used quite differently from how metacognition is being used above. This species has been referred to as ‘procedural metacognition’, ‘experienced-based metacognition’, and ‘evaluative metacognition’, and it is routinely theorized to play some crucial role in explaining the origin of self-knowledge, in particular how it serves as a rudimentary form of mindreading or a precursor to it, one that humans share with preverbal infants and non-human animals (e.g. Smith et al. 1995; Koriat 2000; Hampton 2001; Smith et al. 2003; Mcalfe and Greene 2007; Proust 2007, 2009, 2013; Mcalfe and Son 2012; Bernard et al. 2015; Beran 2019; Smith and Beran 2021). To be clear, I shall refer to this species of metacognition as ‘evaluative metacognition’ because it seems to entail the least conceptual commitment of the others used in the literature, or I shall refer to it simply as ‘metacognition’ when the context makes it clear that I do not mean mindreading.

In this research domain, philosophers and psychologists describe a family of cognitive control mechanisms that function to ensure the adequacy of cognitive operations and the overall fitness of the organism. In other words, here metacognition is a cognitive system often described as having two distinct levels of control, an object level that performs an operation, and a meta-level that produces evaluations of this performance, in the form of feedback signals, which are often referred to in this literature as ‘noetic feelings’, but are also called ‘metacognitive’ or ‘epistemic feelings’ depending on what the particular author believes about the implications these signals have for the overarching cognitive economy and the subject of these feelings herself (see de Sousa 2009; Dokie 2012; Arango-Muñoz 2013; Arango-Muñoz and Michaelian 2014; Proust 2014; Greely 2021). I shall refer to these signals always as ‘noetic feelings’, since it is one of the chief goals of the present work to justify referring to them as ‘metacognitive’ and ‘epistemic’. 
Thus, the present work has the goal of working out how to unify these two species of metacognition, metarepresentational and evaluative metacognition, in light of the empirical evidence from developmental, comparative, cognitive and affective psychology, in order to provide a framework for making inroads into a plausible solution to the ultimate problem of the origin of self-knowledge. In particular, this will be done by showing how noetic feelings, when strongly embodied, prepare cognition and behavior for developing and acquiring capacities for mindreading. But we shall get into the details of this as the present work continues on below. For now, let us return to discussing origin stories of self-knowledge that I think ought to be accepted as starting points.

2. Previously on… The Origin of Self-knowledge

Let us talk about the origin stories about self-knowledge currently on the market that serve as the starting points for the present work. I shall start by introducing a specific telling of the ontogenetic origin story from developmental psychology, the so-called ‘social-scaffolding’ view, that accounts for capacities for mindreading. Thereafter, I shall introduce the phylogenetic origin story told by studies in comparative psychology based on the so-called ‘uncertainty monitoring paradigm’ which accounts for capacities for metacognition. As already stated (see above and Lay Summary), the overarching goal of the present work is to provide a plausible approach to building a bridge between the phylogenetic and ontogenetic origin stories about self-knowledge, while the main target is to solve the problem of triviality facing metacognition.

To begin, notice in the scenario above that, upon responding you are sure about the location of the post office, you have an understanding that, were it to turn out that your belief did not actually match the facts, say, because the red building is, in reality, a coffee shop, and not a post office, then your belief ought to change to match the facts about the world and not, as it were, the other way around, as though the world ought to change to match your belief. In other words, by feeling confident and saying, “Yes, I’m sure”, you are in a mental state made possible by a highly robust species of epistemic agency: you are aware you are a subject that deals in knowledge, beliefs that claim to be justified and true, and you know yourself to be accountable for those beliefs, responsible for ensuring they are justified and true (see Sosa 2015).

At this juncture, one might wish to push back and argue that these notions of responsibility and agency are moral ones (see Schlosser 2015), and it is unclear how exactly epistemic practices
imply ethical commitments. But indeed, the self-knowledge I am interested in explaining is intricately tied to a species of agency that it is as epistemic as it is moral. As people, we hold each other accountable for the beliefs we hold. Beliefs ought to be true and admit of degrees justification which aim, at the very least, to be internally consistent. Should our beliefs be false, we ought to do the right thing and correct them, allow the facts of the world to override what we believe.

If I try to convince you that the red building is the post office, while also sincerely believing it is not, I am lying to you, and you would be right to shun me. Say, I continue to argue with you that this is the post office, say, I do this by misrepresenting the facts, my misconstruing the meanings of my words, twisting them around to meet this false narrative, I am attempting to manipulate you, gaslight you, and, whether I know it or not, cause you to experience cognitive dissonance; and were this about a more serious matter, I might even cause you great psychological harm by doing so. Failing to update one’s beliefs can lead not only to poor epistemic behavior, but also immoral behavior, which is sure to have consequences in any civilized society.

Intertwining self-knowledge with epistemic agency presents an initial conundrum. Our collective epistemic agency is somehow predicated upon an individual’s ability to introspect and consider what she knows. How else could you revise your beliefs and conform to good epistemic practices? But this ability to introspect and know ourselves is also somehow predicated upon our collective epistemic practices. Would you ever learn to monitor and regulate your beliefs as beliefs, that is, as capable of being false and inaccurate, were it not for the fact that your socio-cultural community demands you do, that others will have taught you that regulating your beliefs in this manner is an essential part of being a person, that if you fail to do so, you will be held accountable for your false and inaccurate beliefs in virtue of the fact that they are false and inaccurate?

If it sounds like we are dealing with a chicken-or-egg sort of riddle, it is because we are at this stage. But this is not as problematic as it might first appear, since, of course, we know how to solve the chicken-or-egg riddle, and this solution can point us in the right direction. Basic biology tells us the egg came first, that it was laid by a non-chicken, albeit chicken-like creature. Thus, at the heart of riddles like this lies a kind of conceptual conundrum: there is no way out of the labyrinth so long as the two opposing concepts are held fixed. If Ariadne’s thread is to be found, our stubborn grasp on one of the concepts will need to be relaxed and some flexibility will need to be tolerated. Here the solution is to introduce a chicken-like creature that produced the egg that produced the chicken. Thus, this solution can serve as an analogy for how we ought to go about addressing the conundrum at the heart of the problem of the origin of self-knowledge: let the
empirical research guide the decision about which of the concepts to relax (collective epistemic agency or self-knowledge) and then reframe the question with three concepts instead of just two.

The view I take as one of my starting points is called the ‘social-scaffolding view’, in which self-knowledge emerges as the consequence of specific socio-cultural practices in the human enculturation process (see Vygotsky 1930-1934/1978; Gopnik and Wellman 1993; Heyes 1993, 1998, 2014, 2018; Clark 1994; Heyes and Bennet 1996; Hutto 2008; Zawidzki 2013; Heyes and Frith 2014; Shea et al. 2014; Fenici 2017). Evidence for the social-scaffolding view with respect to self-knowledge comes largely from developmental psychology and cross-cultural analysis. Skills related to self-knowledge arrive late in development, and at which point, and to which degree, will depend upon how skills surrounding belief attribution are taught and cultivated within an individual’s socio-cultural environment (see Heyes et al. 2020). If the culture is such that belief-state ascriptions are highly prevalent, then skills related to self-knowledge will emerge earlier and be more finely tuned to reflect accuracy than in cultures where belief-state ascriptions are less prevalent. Should those practices be somehow absent in an individual’s development, this individual will fail to acquire capacities for self-knowledge.

Though not entirely uncontroversial, the social-scaffolding view continues to garner support, and I shall assume it to be on the right track, so that socio-cultural practices forging a collective epistemic agency came before capacities for acquiring self-knowledge (and not the other way around). In keeping with our analogy then, the order of steppingstones on the path to self-knowledge ought to be construed like this: first came a non-self-knowledge albeit self-knowledge-like state, which, in turn, produced social-cultural practices, which cultivated a collective epistemic agency, which, in turn, made states of self-knowledge possible.

So, with the answer to the question about the ontogenetic origin of self-knowledge being the socio-cultural practices that cultivate a collective epistemic agency, the question becomes whether these socio-cultural practices have their origin in internal states of the individual members comprising this on-the-path-to-becoming epistemic community. Similar to how some non-chicken

---

3 The major contender to the social-scaffolding view is the nativist view, in which the mindreading capacities are not acquired through development and enculturation but are rather the products of natural evolution and are thus native (intrinsic, innate, etc.) capacities to distinctly-human cognition. Since the present work takes the social-scaffolding view as its starting point, reasons for and against the nativist view will not be discussed here. For more on this view, see Carruthers and Smith (1996), and for debates between nativist and proponents of the social scaffolding view, see Fenici and Zawidzki (2021).

4 Though it might be somewhat banal as far as analogies go, let us continue to revisit the chicken-or-egg analogy if only to keep the three crucial components separate and in the right order: (i) the self-knowledge-like state (or the chicken-like creature), which produces (ii) the social-cultural practices (or the chicken-egg), which, in turn, forge a collective epistemic agency and make (iii) self-knowledge possible (or the chicken itself).
albeit chicken-like creature produced the egg that produced the chicken, the crucial socio-cultural practices might have originated from some non-self-knowledge albeit self-knowledge-like state.

This is precisely what the present work endeavors to account for. Ultimately, it will be argued that a family of cognitive control mechanisms, ‘embodied metacognition’, produces strongly embodied noetic feelings, crucial enablers and facilitators of socio-cultural practices that lay the foundation for a collective epistemic agency that makes self-knowledge possible (see Chapters 2, 4, and 6; Section 4 below). Thus, the overarching task set by the present work is providing a plausible account of how specific internal states, those characterized by strongly embodied noetic feelings, are crucial ingredients to self-knowledge, though the main target is to settle the dispute between metarepresentationalists and evaluativists about the relevancy of metacognition (i.e. solve the problem of triviality).

To be clear, the chicken-or-egg riddle is not meant to be an argument, but rather an analogy for the general gist of a prominent view already present in the literature (most notably in Proust 2013), one which the present work uses as a starting point. However, unlike the social-scaffolding view, this view is not assumed by the present work to be correct; rather, I shall endeavor to demonstrate that the appeal to noetic feelings is on the right track for solving the problem of triviality, on the one hand, and explaining the origin of self-knowledge, on the other.

Evidence that noetic feelings serve as crucial precursors to self-knowledge comes largely from studies in comparative psychology, in particular those that employ the so-called ‘uncertainty monitoring paradigm’ (e.g. Smith et al. 2003; Smith and Washburn 2005; Smith et al. 2008; Beran et al. 2014; Zakrzewski et al. 2018; Beran 2019), in which various species of animal are presented with tasks designed to measure their capacity to monitor and mitigate uncertainty. Their performance is then compared to that of humans to determine whether their abilities are on par. Non-human animals with comparable abilities are then argued to possesses a phylogenetic precursor to self-knowledge, specifically a kind of self-confidence theorized to be mediated by noetic feelings (for more on the details of this paradigm, see below Section 4 as well as Chapters 1 and 5). Applying our analogy thus yields the proposal that noetic feelings of confidence are the chicken-like creature that produced the egg that produced the chicken.

Notice, of course, that just because it was the egg that came first, it does not mean the chicken-like creature is somehow insignificant to the natural origin story of the chicken. Similarly, the self-knowledge-like state that produces the socio-cultural practices ought to be construed as crucial to the origin of self-knowledge, so long as an account can be provided that justifies construing this proto state as such, rather than, say, a state that in no way resembles self-knowledge,
but is nonetheless essential to the story (e.g. cognitive states enabling adaptive behavior). Surely, the correct response is to say that the self-knowledge-like state is crucial in virtue of being the state that made those socio-cultural practices possible, which, in turn, made self-knowledge possible. In other words, I understand cruciality to an origin story to consist in either being the origin itself or providing the sufficient enabling conditions of the origin, and, as such, the overarching argument in the present work will follow this line of reasoning. That is, I shall endeavor to make a plausible case for construing noetic feelings produced by embodied metacognition as crucial to the origin story of self-knowledge in the above sense, specifically in virtue of having plausibly laid the foundation for the socio-cultural practices that make self-knowledge possible.


As already introduced, the self-knowledge-like states that lend themselves to being construed as significant precursors to the socio-cultural practices that comprise a collective epistemic agency and thus lay the groundwork for full-blown states of self-knowledge, are a peculiar species of mental entity called ‘noetic feelings’. At their most cognitively and psychologically basic, noetic feelings are important ingredients in cybernetic theories, serving as feedback signals produced by cognitive control mechanisms for the purpose of guiding both covert cognitive processes and overt behavior to conform to descriptions set by cognitive models of internal processing standards relative to those processes and behaviors. The ultimate result of a system so equipped are highly flexible behaviors that aim at accuracy. Already mentioned above (Section 2), these cognitive control mechanisms are often referred to collectively as ‘evaluative metacognition’.

There is a manner in which noetic feelings are clearly about cognition, and, as such, are not too terribly divorced from the complex metabeliefs of metarepresentational metacognition, beliefs about beliefs. But to make sense of how noetic feelings might serve as potential precursors to metabeliefs, as well as the problems facing this proposal, it will help to first grasp how noetic feelings are incommensurate with metabeliefs. To arrive at this, let us discuss details surrounding the role that noetic feelings are theorized to play in the overarching cognitive economy.

I find it helpful to construe noetic feelings of confidence (and their inverse, uncertainty) as the most basic and essential of all noetic feelings. When feeling confident in your abilities, you tend to follow through on them, and when feeling uncertain, you tend to become insecure, reluctant,
and not engage in action. The creature that possesses well-calibrated noetic feelings of confidence (that also heeds these feelings) is one that is highly inclined to act when most advantageous and demotivated to act when least advantageous. The upshot of well-calibrated noetic feelings are flexible responses that have been fine-tuned by evolution and development toward performing those actions well, that is, accurately, according to some standard of performance set by both nature and the individual’s learning history.

Recall that the form of self-knowledge I am interested in explaining requires the capacity to entertain beliefs as beliefs, as mental entities that can fail to accurately represent the world. It is thus this implicit notion of accuracy at work in these control mechanisms, as well as this implicit demand to conform behavior to accuracy, that has been theorized to justify construing noetic feelings as crucial precursors to a collective epistemic agency, and in virtue of this link, significant precursors to self-knowledge (see Proust 2013; for more on Proust’s view, see Chapter 1). Noetic feelings are thus the products of a species of metacognition that is theorized to serve as a crucial phylogenetic precursor to mindreading.

At this point, it might be best to introduce you to a long-standing debate in the metacognition literature that has resulted in a stalemate. It is the chief aim of the present work to solve, or making inroads into solving, the various problems situated at the heart of this debate, in particular the problem of triviality (for more on these problems, see Section 4). If the arguments made by the present work are sound, specifically those made in Chapters 2, 4, and 6, then this stalemate will finally be at an end.

On one side of the debate stands proponents of the view that metacognition, in virtue of noetic feelings, ought to be construed as a crucial precursor to self-knowledge. But opponents have argued that it is deeply unclear how metacognition can fulfill this role since it is unclear how metacognition is anything over and above mere first-order cognition (see Carruthers 2008; Carruthers 2009a, 2009b, 2014, 2016; Carruthers and Williams 2019; Nicholson et al. 2019). In other words, opponents demand that something equivalent to, or conceptually on par with, the

---

5 ‘Implicit’ and ‘explicit’ are common terms in the literature on metacognition for making sense of how the two families of metacognition are distinct, the evaluative sort, which provides implicit evaluations of cognitive performance (or feelings), versus the metarepresentational sort, which produces explicit evaluations (or judgments; see Shea et al. 2014). What is often meant by this distinction is the degree to which the subject herself is aware of what the metacognitive evaluation entails or implies. Explicit evaluations can be thought of as exclusionary, since all but those subjects that possess the cognitive capacities for understanding the implications of the evaluation are excluded from dealing in them. Meanwhile, implicit evaluations are the inverse, inclusionary, since even subjects that do not possess the cognitive capacities for grasping the implications of the evaluation are included in the set of subjects that deal in them. So in this case, accuracy is implicitly implicated in these cognitive evaluations of performance, as the subject need not be in the position to understand that her performance is aiming at accuracy (and all this implies) in order to be guided by the evaluation and, crucially, for us to explain the role of the evaluation in the overarching cognitive process.
second-order structure of metabeliefs, beliefs about beliefs, be somehow reflected in the products of metacognition in order that it be construed as a crucial precursor to self-knowledge. This might sound like one problem, but it is actually two problems. On the one hand, there is the problem that metacognition might not amount to some degree of cognitive sophistication over and above that of mere unsophisticated, first-order cognition. Meanwhile, on the other hand, there is the problem that metacognition might not be a crucial ingredient for the origin of self-knowledge.

Providing an account of metacognition that makes it clear how it is not ‘run-of-the-mill’, first-order cognition, that there is something significant about it that makes it over and above the day-to-day business of cognition, is a deep conceptual problem in the debate over the role of evaluative metacognition. Solving it amounts to solving what I shall call ‘the problem of triviality’, since, if it is not solved, it is unclear how evaluative metacognition is particularly crucial in terms of producing sophisticated cognitive behaviors involved in reflective action (for more on this problem, see Chapter 2; for plausible solutions to it, see Chapters 3 and 5). On top of this, solving the problem of triviality in such a manner that it is clear, not only how metacognition is not merely run-of-the-mill cognition, but also, crucially, a plausible precursor to the socio-cultural practices, where from self-knowledge originates, amounts to solving the problem of triviality in such a way that also the ultimate problem is solved, ‘the problem of the origin of self-knowledge’.

Do take note that the chief goal of the present work is to solve the problem of triviality, and thereby end the stalemate that has plagued the literature on metacognition for decades. For this problem, I shall offer two possible solutions. The first is embodied metacognition, argued for here, specifically in Chapters 2, 4, and 6. The second solution is a non-embodied but nonetheless conscious and executive form of metacognition presented in Chapter 5 but argued in Chapter 6 to be less preferable to embodied metacognition for various reasons. That is, after providing two solutions to the problem of triviality, I shall endeavor to draft up the blueprints for building a bridge between the phylogenetic and ontogenetic origin stories about self-knowledge in Chapter 6, and thus providing an account of how to unify metacognition with mindreading.

This distinction between metacognition and mindreading, as well as the potential for metacognition to serve as its precursor and thus serve as a precursor to self-knowledge, will be the chief subject matter of the present work, and so it will be highly beneficial to draft up a sketch of all this now, the controversy, as well as the possible solution-space. But to motivate the plausibility of unifying the two, it will first be helpful to deploy more technical notions for talking about metacognition as a whole, so let us dive straight into it; afterward, challenges facing this proposal will be introduced (see Section 4).
Recently, mindreading has been called ‘system 2 metacognition’ in order to distinguish it from evaluative metacognition, which has been dubbed ‘system 1 metacognition’ (e.g. Fleming 2021). This manner of distinguishing the two is based on one of the most influential theories in cognitive psychology, the systems theory of cognition, which, as the name implies, describes cognition as effectively split into two distinct systems, each serving as sort of the inverse of the other (Kahneman and Tversky 1982; Kahneman and Frederick 2002; Kahneman 2011).

Consider calculating a sum. Say, you have just finished eating dinner at a restaurant and are now in the process of determining how much tip to leave the server. Here you would likely do a quick calculation in your head: “It may not be exactly 15%... but it’s close enough”. Contrast this to a similar task, but one that demands accuracy, such as doing your taxes. Likely, you will want a calculator to offload the demands of this task onto an external device. Why would you want to do this? It’s obvious, of course. Doing so will not only help mitigate potential mistakes in calculation, but it will also save time and effort, since doing these sums in your head, while also aiming at accuracy, would only make the task more effortful and time-consuming.

But notice how much slower and more deliberate this latter task would be relative to the first task in which you made a rough guess. Such differences are characteristic of the distinct systems of cognition. System 1 is the fast and efficient, often intuition-based system, in which most of the cognitive processing is performed below the threshold of conscious awareness, while system 2 is the slow and time-consuming system that depends upon resources in working memory and whose cognitive processing is often performed consciously and deliberately.

Studies on mindreading have demonstrated how the entertaining of metabeliefs is highly dependent on the availability of working memory, requiring the online control of thought, and so the mechanism that produces metabeliefs is theorized to be a system 2 sort of mechanism (see Carruthers 2017). Meanwhile, metacognition is capable of producing feedback signals, noetic feelings, without thereby placing strong demands on working memory, and so the mechanism that produces noetic feelings is believed to be a system 1 sort of mechanism (see Proust 2010).

Not only does the systems theory of cognition provides us with a useful framework for thinking about how metacognition and mindreading are distinct, but it also assists us in grasping how the latter (mindreading) might actually depend upon the former (metacognition). Consider once again the scenario above with the metabelief about the location of the post office. Concentrate

---

6 Though Kahneman (2011) cautions against construing each system as a particular mechanism or family of mechanisms, it is common in the metacognition literature to refer to distinct forms of metacognition as underpinned by distinct mechanisms, such that either mechanism shares a family resemblance with one of the two systems (see Heyes et al. 2020).
on the basis upon which you made the decision that your belief was true, i.e. what motivated saying, “Yes, I’m sure”. One construal of this depicts you performing a mental action, a judgment. You entertained the belief about the location of the post office and judged this belief to be true. But on what basis did you make this judgment? Did you take the time to consider the evidence one way or the other? – obviously not, since you do not need to: a snap judgment suffices.

So now the scenario above, which we took to illustrate mindreading capacities, in which you turn mindreading capacities upon yourself, appears to illustrate a process that depends upon system 1 mechanisms that enable capacities for metacognition. The metabelief that the belief about the location of the post office is true is produced via a mental action, a judgment, made on the basis of a rapid and almost effortless evaluation of the evidence, an evaluation produced by system 1 mechanisms enabling capacities for metacognition.

To be clear, one would not claim that the judgement was produced by the metacognitive mechanism, but rather that the feeling of confidence, which shapes the judgement, was so produced. Thus, the pivotal question becomes whether you would have cast the judgment that the belief was true had you not felt confident it was. If one is inclined to think that the judgement is a mental action and, as such, requires some positive attitude or motivation in order that it be performed (see O’Brien and Soteriou 2009), then one will be sympathetic to this manner of establishing a link between metacognition and mindreading, such that noetic feelings facilitate mindreading.

But notice that one does not need to believe that a feeling of confidence is necessarily implicated in the casting of judgment in order to believe that the feeling is often a strong motivator for casting judgements. From this more conservative view it also follows that metacognition guides and facilitates mindreading. In order to also accept the proposal offered by the present work about how to bridge mindreading and metacognition, and thus unite the ontogenetic and phylogenetic origin stories about self-knowledge, one need only accept this conservative view: noetic feelings can serve to motivate and guide reasoning and decision-making and thus facilitate mindreading.

Thus, metacognition is a species of cognitive control that produces evaluations, feedback signals in the form of noetic feelings, that serve to guide our mental and epistemic actions, theorized by proponents of metacognition to be the source of our collective epistemic agency. To be clear, this view is precisely what will need to be made plausible by the present dissertation in light of its various criticisms, but the following example should serve as initial motivation for thinking about how this is supposed to work.
If I try to convince you that the coffee shop across the street is actually the post office, you
do not need to weigh up the evidence in order to judge whether what I am claiming is true. Instead,
you quickly and effortlessly get a sense (a feeling) that I am (frankly) just bullshitting you (see
Frankfurt 2005). This sense, this feeling-based evaluation, thus guides and informs your epistemic
practice of calling me out as a liar. Indeed, noetic feelings are often cited as evidence in explaining
the decision to hold someone morally and epistemically accountable: “It felt like he was gaslighting
me!”. Quite plausibly then, certain episodes of cognitive dissonance are intertwined with noetic
feelings of uncertainty and confidence. So, the story goes, noetic feelings and evaluative
metacognition reflect the specific cognitive conditions that make the socio-cultural practices that
cultivate our collective epistemic agency possible and thus ought to be construed as crucial
precursors to self-knowledge, a proposal that will be developed throughout the present work (in
particular, in Chapters 2, 4, and 6).

4. We Need to Talk About Kevin: or The Problem(s) with Noetic Feelings

That being said, there are issues with this proposal. For one, the mechanisms that underpin
capacities for metacognition, as commonly construed, are not too terribly sophisticated and so it is
unclear how these mechanisms are anything above and beyond those mechanism that enable basic
feats of cognition (for more on this, see Chapters 3 and 5). Mechanism for metacognition are
often described in the vernacular of cybernetics, largely borrowed from early work in motor
cognition, in which two distinct control structures are described as making reinforcement learning
possible, one primary structure responsible for performing the action and another, meta-structure,
responsible for monitoring the performance of the action and producing feedback. These same
control structures are used to explain basic feats of motor cognition, and so, at first glance, it is
unclear how any of this is complex enough to warrant construing these mechanisms for
metacognition as sophisticated advancements over and above basic cognition, let alone significant
precursors to self-knowledge.

To refer back to our analogy, it is deeply unclear how metacognition, as it has been so far
articulated, is chicken-like enough to be the creature that produced the egg that produced the
chicken. Rather, metacognition might instead be a family of mechanisms that produced some
adaptive behavior that make the way for other forms of cognitive control, and so on and so on,
down the line, such that metacognition does not reflect such a crucial precursor to self-knowledge
after all. In this respect, the solution is thus to provide an account of the self-knowledge-like state that plausibly produced those socio-cultural practices that make self-knowledge possible.

As already discussed, the present work endeavors to show that noetic feelings are the plausible candidates, serving as the exact precursors. It will emerge below, however, that not just any species of noetic feeling can serve this role. I argue that only noetic feelings produced by an intricate brain-body interface can serve as the precursor we need. In this respect, let us consider an illustration of the issues facing the appeal to noetic feelings to shore up metacognition against the problems of triviality and the origin of self-knowledge. This will not only help us make sense of what is problematic about how noetic feelings as currently described, but it will also motivate strongly embodied noetic feelings as the possible solution.

Imagine George, a tree-dwelling spider monkey exploring his arboreal habitat, swinging from tree limb to tree limb, when he is suddenly uncertain whether a tree limb ahead is within his reach. That is, in some sense that does not need to be specified as of yet, it seems reasonable to say that George will become uncertain about his capacity to swing to the tree limb (for this example, it does not make a difference whether this uncertainty is conscious or unconscious). As a result of this uncertainty, George will stop his joyful swinging and instead look for another tree limb to swing to, or perhaps something else to occupy his time. But how did George make this decision?

Adopting any cybernetic theory of action, theories which continue to serve as the bedrock of contemporary cognitive psychology, George’s behavior, that which is associated with his becoming uncertain, is explained by the production of a feedback signal, one that is produced by mechanisms for cognitive control, a signal which informs the system about the inadequacy of behavior to conform to internal models describing a standard of performance respective of the particular behavior, models which have been honed by both evolution and George’s own learning history. The question now becomes whether this feedback signal is a noetic feeling – because, if this is all it takes to be a noetic feeling, then proponents of metacognition have a serious issue on their hands, since it is deeply unclear how such cognitively basic feedback signals are anything other than the signals produced by mechanisms for adaptive behavior, called Reward Prediction Error signals or RPEs (see Chapter 3), let alone crucial to producing the socio-cultural practices that cultivate a collective epistemic agency and thus make self-knowledge possible. Distinguishing mechanisms for metacognition from more basic mechanisms for adaptive behavior is the chief goal and first major challenge of the present work, i.e. solving the problem of triviality.

As made clear by this illustration, it is unclear how noetic feelings are little more than signals theorized to underpin basic motor cognition and so cannot be in any way distinctive of
human cognition and, perhaps, that of a small set of non-human animals. Indeed, these signals seem to be little more than those which explain the day-to-day business of cognition, and so there is little in this analysis to support the proposal that these signals will produce sophisticated cognitive behaviors – let alone, give rise to capacities for self-knowledge – so that noetic feelings would need to be recognized as distinct from signals that produce basic feats of cognition. In other words, failing to give an account of these signals, and the mechanism that produces them, as something over and above mere run-of-the-mill cognition is to render evaluative metacognition trivial, and thus fail to address the problem of triviality introduced above.

In what follows, this problem will be explored in detail as it faces an account of noetic feelings. Essentially, the mechanism that has been theorized to produce noetic feelings tends to be described as so cognitively basic that metacognition, as of yet, cannot plausibly reflect a significant precursor to self-knowledge. This problem will be made apparent by showing the conceptual proximity of metacognition to models of reinforcement learning (see Chapter 3), which do not require any reflective or deliberative capacity, as well as its close conceptual proximity to the systems of feedback described by the prominent theory of predictive processing (see Chapter 5), systems which are described as operative at practically every level of the cognitive economy.

The solution to the problem of triviality presented and defended by this work involves demonstrating that mechanisms for metacognition are not only more sophisticated than those underpinning reinforcement learning and error dynamics in predictive processing, but so sophisticated so as to be an enabler of robust forms of self-control, conscious control over cognition and behavior (see Chapter 4). Regarding the mechanism itself, this is argued to comprise an intricate brain-body interface, in which cerebrally-generated feedback signals, noetic feelings, become intertwined with extracerebral (bodily) changes perceived by the subject through interoceptive channels, in particular bodily changes produced by the cardiovascular system, and hence the subtitle of the dissertation: “How We Feel Our Hearts to Know Our Minds” (see Chapter 2).

Thus, the solution to the problem of triviality offered by the present work, one that has eluded philosophers and psychologists for decades, amounts to appealing to noetic feelings as strongly embodied, as conscious expressions of emotion produced by an intricate brain-body interface, which, due to how neurocomputational changes are represented in the de re mode of reference, enables subjects to exercise conscious control over cognition and behavior. Such capacities are, of course, by no means trivial, and reflect highly sophisticated form of self-control, thus warranting the title ‘metacognition’ in the name ‘embodied metacognition’ (see Chapter 4).
Before moving on to sketch the framework for making inroads into the problem of the origin of self-knowledge proposed by the present dissertation, allow me to be explicit about what I take consciousness to consist in. Here I wish to refer to what has been called ‘access consciousness’ (Block 1995), which Shea and colleagues (2014) describe in terms of a mental representation being “access conscious just in case it can be used, without further processing, for verbal report, inferential reasoning, storage in episodic and semantic memory, and by other consuming systems” (186). Specifically, conscious cognitive processing, in the sense deployed here, involves access to information at the personal (or animal) level of description for the control of cognition and behavior enabled by the executive control network (see Gazzaniga 1988; Perner and Lang 1999; Dehaene 2001; Perner and Dienes 2003; Dehaene and Changeux 2004).

Now, in order to arrive at a plausible framework for making inroads into solving the ultimate problem of the origin of self-knowledge, the present account of metacognition, embodied metacognition, will also need to ensure these added layers of sophistication are such that they explain the enabling conditions of those socio-cultural practices that cultivate a collective epistemic agency that, in turn, makes self-knowledge possible. To better introduce how this will be argued for, let us discuss two corollary problems that will need to be solved before we can pivot away from the solution to the problem of triviality and toward confronting the problem of the origin of self-knowledge.

The problem of triviality can be made salient by two related problems: ‘the problem of metacognitive biodiversity’ and ‘the problem of honeybee metacognition’. These problems emerge out of the empirical research that comprises the foundation upon which claims receive justification about metacognition serving as a phylogenetic precursor to self-knowledge. As such, these two problems will need to be adequately addressed before any inroads can be made into solving the problem of the origin of self-knowledge.

Already mentioned above (Section 2), this evidence comes largely from the uncertainty monitoring paradigm in comparative psychology. Similar to what is going on in the example with George above, the uncertainty monitoring paradigm presents various species of animal with tasks designed to measure their capacity to monitor and mitigate uncertainty. Their performance in these tasks is then compared to that of humans to determine whether their ability to monitor and mitigate uncertainty is on par with our abilities. Though each experiment has its particular nuances, the basic setup remains the same. The subject (human and non-human animal alike) is confronted with a battery of perceptual discrimination tasks, such as determining whether the pitch of an auditory signal is either high or low, with a subset of tasks designed to be extremely difficult (or even
impossible) for the subject to categorize correctly. In these hard-to-discriminate trials, the subject is at chance and their performance will decrease relative to easier trials.

The fascinating aspect of this paradigm is how researchers measure the difference in performance during these difficult trials before and after the introduction of an opt-out mechanism (often a button, pedal, or an icon on a computer monitor) that the subject must learn to properly engage with in order to skip hard trials and prevent the punishment associated with failing to discriminate. The question becomes whether the animal will learn to use the opt-out button so as to improve its performance on par with how humans can. For if the animal can do so, many researchers in comparative psychology believe we should say, this implies the animal possesses some sense of when it knows versus when it does not, that is, possesses some rudimentary form of self-knowledge serving as a phylogenetic precursor to our sophisticated form.

Though skeptics have raised various issues with this paradigm regarding whether the animal is simply learning to use the opt-out button by mere associative learning (Le Pelley 2012, 2014), or whether the animal’s use of the opt-out button is to be explained by the mapping of a ‘middle option’ rather than an ‘I don’t know option’, this paradigm has gone through dozens of revisions over the span of three decades to attempt to address these concerns, and it has come to change the minds of some skeptics, though holdouts exist (see Chapter 5 for more). Without a doubt though, this paradigm provides exciting results about metacognition, since it routinely shows capacities for monitoring and mitigating uncertainty are not equally distributed across the animal kingdom.

This unequal distribution raises the problem of how to make sense of this diversity in metacognitive performance, and hence the problem of metacognitive biodiversity, which no account of metacognition has set itself the task of explaining. By the end of the present work, I will have provided a highly nuanced theory of metacognition, one that distinguishes subpersonal from personal kinds of confidence, as well as non-feeling-based and feeling-based kinds of confidence that will help us make sense of how to address the problem of metacognitive biodiversity in such a manner that the problem of triviality is also resolved (see Chapter 6; and for an alternative solution, see Chapter 5).

One fascinating case that has the potential to make clear how the problem of triviality is wrapped up with the problem of metacognitive biodiversity, and how these problems culminate in the problem of the origin of self-knowledge, is the case of metacognition in honeybees. Perry and Barron (2013) conducted an influential study based on the uncertainty monitoring paradigm in honeybees and found that their metacognitive performance is quite nearly on par with that of mammals, primates, and humans. This raises serious doubts about the possibility of metacognition
to comprise sophisticated cognitive machinery that could potentially serve as a crucial phylogenetic precursor to advanced form of self-knowledge, simply due to the fact that over 600 million year of evolution separate the hominid line from the apidae line of the honeybee. The problem of honeybee metacognition demands a solution, I think, that does justice to their impressive metacognitive capacities while also securing the species of metacognition in the mammal branch of life as a crucial precursor to collective epistemic agency and advanced forms of self-knowledge (see Chapters 5 and 6).

The present work discusses two distinct accounts that offer a solution to these three problems (i.e. the problem of triviality, the problem of metacognitive biodiversity, and the problem of honeybee metacognition). On the one hand, there is the proposal of an embodied metacognitive mechanism, an intricate brain-body interface that produces strongly embodied noetic feelings, conscious expressions of emotion. This, of course, is the proposal put forward by the present work (see Chapters 2 and 4). Meanwhile, an alternative proposal offered by Meyniel and colleagues (2015a, 2015b) is discussed in Chapter 5. Though this account also appeals to metacognitive feedback signals as conscious to solve the problems facing metacognition, the mechanism that produces these signals is described as solely cerebral and does not depend upon extracerebral processes.

That said, as the title of the present work suggests, it will be made clear why an embodied account of metacognition is the preferred path to solving these problems, as well as potentially offering the right way forward in addressing the ultimate problem of the origin of self-knowledge. With regard to solving the three initial problems, several empirical studies at the intersection of metamemory and interoception (i.e. the internal perception of extracerebral changes) will be discussed that demonstrate the remarkable degree to which capacities for metacognition are enabled and facilitated by extracerebral processes (see Chapter 2 and 3). This evidence thus supports an embodied over a non-embodied account of conscious metacognitive feedbacks signals (for the detailed argument, see Chapter 6). What is potentially more, it will be argued that an embodied account is better equipped to provide us with the necessary tools for building the bridge between the phylogenetic and the ontogenetic origin stories about self-knowledge (also Chapter 6). So, let us turn to a sketch of this now.

5. The Sketch: How Feelings our Hearts Enables us to Know our Minds: or Strongly Embodied Noetic Feelings
The solution offered by the present work to these three problems introduced above amounts to appealing to noetic feelings as strongly embodied. Not only does this make the mechanism that enables capacities for metacognition sufficiently robust, so that we would surely think twice about attributing it to all manner of animal, but more importantly it helps illustrate what might have been the exact conditions under which socio-cultural practices for cultivating collective epistemic agency emerged. The case for embodied metacognition as offering a plausible framework for solving the problem of the origin of self-knowledge is a three-pronged argument.

First, due to how strongly embodied noetic feelings are embodied states with strong informational correlations to neurocomputational changes, which, in turn, describe the adequacy of cognitive acts in conforming to dynamic profiles of performance, I argue, these states ought to be construed as candidate states for precursors to self-belief states. Put simply: the first proto self-belief state that nature ever produced (at least in our branch of the tree of life) was the state of confidence and its inverse, uncertainty (see Chapter 4).

Next, due to their status as conscious expressions of emotion, these robust psychological states are in a privileged position to produce rich epistemic practices that enable a species of intraindividual regulation that approximates the updating of beliefs in light of evidence and counter evidence (see also Chapter 4). As feelings, noetic feelings are emotional states that, due to their nature as either positively or negatively valenced with some degree of bodily arousal, intrinsically motivate or demotivate to some corresponding degree. Crucially, feelings ought to be construed as imperatives, demanding subjects resolve them in an emotionally satisfying manner.

If feelings are negatively valenced, they will demand to be replaced by positively valenced feelings, motivating subjects to transform noetic feelings of uncertainty into noetic feelings of confidence. Meanwhile, if noetic feelings are positively valenced, this will motivate subjects to explore the limits of her noetic feelings and thus explore the epistemic opportunities afforded by her environment. Indeed, evidence of animals engaging in this behavior is already observed in those animals that do well in the uncertainty monitoring paradigm (e.g. Smith et al. 1995; Hampton 2002). Once feelings are finely tuned to index neurocomputational changes that describe the adequacy of cognitive performance, the result is thus rich epistemic practices in the individual for exploring, exploiting, and making sense of their environments and the practices themselves (this is discussed in Chapters 4 and 6 in detail).

Finally, as strongly embodied, noetic feelings can readily transform into overt bodily acts, expressions, gestures, and reactions. This means, noetic feelings can be produced in an externally
observable manner so as to be detected, monitored, and regulated by other individuals in the socio-cultural environment, and thus enable, not only a species of intraindividual but also a species of interindivdual regulation. Consider all the multifaceted ways that confidence can be expressed with the body. Not only are there particular facial expressions that we have learned to associate with confidence and the lack thereof, but there are also specific hand gestures, specific postures, specific gaits. Just think of the school child who quickly and vigorously raises her hand whenever the teacher asks the class a question. Everyone can see just how confident she is that she believes she has the correct answer. The present work thus hypothesizes that behaviors such as these are the foundation upon which socio-cultural practices for cultivating collective epistemic agency are built.

A species that learns to detect bodily expressions of noetic feelings and learns to use them as a basis for decision-making, for example in determining social-role assignments, will be one that will incur a massive survival advantage because, when well placed, confidence approximates competence. Of course, noetic feelings of confidence tend to be well placed, since these feelings are produced by mechanisms that have been finely tuned by nature and the individual’s learning history to ensure survival. For these reasons then, strongly embodied noetic feelings ought to be construed as significant precursors to self-knowledge, since it is in virtue of them that the conditions arose in which, our socio-cultural practices for cultivating our collective epistemic agency emerged. In the end, directions for future research and testable empirical predictions that emerge from this proposal will be discussed (see Conclusions).
Summary of Chapters

Chapter 1. What is Evaluative Metacognition? Proust’s Cybernetic Theory of the Animal Mind

The purpose of Chapter 1 is to provide a thorough accounting of Proust’s philosophy of mind with the aim of explaining her theory of metacognition, perhaps the most fully articulated account of evaluative metacognition in the literature. With respect to addressing the problems facing metacognition, it will be demonstrated how Proust’s case rests on arguing for how mechanisms for metacognition explain the origin of a normativity of an epistemic kind, which, in turn, paves the way for a subject-world distinction in thought, distinctive of human cognition. However, despite the sophistication of epistemic normativity, this account of what makes metacognition special will be demonstrated in Chapter 3 to struggle to solve the problem of triviality.

Chapter 2. Are the Noetic Feelings Associated with Semantic Memory Embodied? The Case for Embodied Metamnemonic Cognition

After having discussed the main account of evaluative metacognition above, Chapter 2 presents a discussion of the evidence upon which the case for embodied metacognition, an augmentation of evaluative metacognition, will be built in Chapter 4. Here the discussion centers around how best to account for the mechanism that produces noetic feelings in light of recent empirical evidence from interoception-based studies in metacognition. The result is a Neo-Jamesian theory about noetic feelings: what one feels, whenever undergoing a noetic feeling, are bodily changes, sensed through interoceptive channels. Finally, mental states characterized by noetic feelings are argued to represent (in the de re mode of reference) neurocomputational changes.

In Chapter 3, the literature on metacognition is revisited and the aim is to expose problems facing two approaches to shoring up metacognition against the problem of triviality. One approach is inspired by Shea’s peculiar account of metarepresenting, in which subpersonal representations are construed as metarepresentational if their content is about the content of another subpersonal representation. This chapter argues that if this approach is taken, it will ultimately fail because, in order for such an approach to respond to criticism, it will need to appeal to conceptual capacities, the origin of which metacognition seeks to account for.

The second approach discussed is Proust’s cognitive role approach already laid out above in Chapter 1. However, in Chapter 3 this approach is demonstrated to encounter difficulties when attempting to solve the problem of triviality due to how it relies upon a model that shares a strong family resemblance with a model for reinforcement learning. As a result, this approach struggles to articulate sufficiently sophisticated cognitive architecture required to solve the problem of triviality. Finally, through contrasting noetic feelings with reward prediction errors signals in reinforcement learning, the proposal arises that construing noetic feelings as conscious could potentially solve the problem of triviality, a proposal pursued in the next chapter.

Chapter 4. The Case for Embodied Metacognition: Shoring up Metacognition against the Problem of Triviality

By appealing to the empirical evidence discussed in Chapter 2, Chapter 4 makes the case for embodied metacognition, an augmented version of evaluative metacognition, in which the crucial metacognitive feedback signals, noetic feelings, are construed as strongly embodied, conscious expressions of emotion, characterizing states of executive control over cognition and behavior. The appeal to consciousness and executive control is argued to solve the problem of triviality, the chief goal of the present dissertation. Furthermore, it is argued that these properties suggest mental states characterized by noetic feelings are proto states of self-belief. Thus, mental states characterized by noetic feelings serve as one of the crucial planks for building the bridge between the phylogenetic and ontogenetic origin stories argued for in Chapter 6.
Chapter 5. Evaluative Metacognition and the Predictive Processing Framework: Grappling with the Origin of Self-knowledge

The purpose of Chapter 5 is to discuss how best to integrate metacognition into the fold of the predictive processing framework. In so doing, a proposal by Meyniel and colleagues will be discussed as an alternative to the one offered by the present work and argued to solve the problem of triviality and its related problems (that of metacognitive biodiversity and of honeybee metacognition) in a similar manner, namely by appealing to conscious metacognitive feedback signals. Distinguishing Meyniel and colleagues proposal from the one offered here is how metacognitive feedback signals are described in their model as non-embodied, solely central and cerebral feedback signals characterizing states of confidence. However, it will be demonstrated how this theory does not account for the recent empirical evidence from interoception-based studies on metacognition, nor does it offer a clear framework for making inroads into the overarching problem of the origin of self-knowledge.

Chapter 6. How Feeling Our Hearts Enables Us to Feel Our Minds: Uniting the Phylogenetic and Ontogenetic Stories about the Origin of Self-knowledge

The final chapter, Chapter 6, aims to demonstrate why the case for embodied metacognition, with its appeal to strongly embodied noetic feelings, ought to be preferred to the alternative account offered by Meyniel and colleagues. Meyniel and colleagues’ account is first demonstrated to be easily augmented to describe an embodied metacognitive mechanism and then reasons are provided for why this augmented account (effectively, embodied metacognition) ought to be preferred. On the one hand, it is better able to explain the empirical evidence about the intricate brain-body relationship in the production of noetic feelings discussed in Chapters 2 and 4, and, on the other hand, it is better suited for making inroads into solving the problem of the origin of self-knowledge. This is because of how strongly embodied noetic feelings assist us in accounting for both intraindividual and interindividually regulatory practices of an epistemic kind that could serve as the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge.
Chapter 1

What is Evaluative Metacognition? Proust’s Cybernetic Theory of the Animal Mind

Abstract

The ultimate goal of this dissertation is to establish the family of mechanisms and capacities known as ‘evaluative metacognition’ as a plausible phylogenetic precursor to capacities for self-knowledge (henceforth, ‘metacognition’). In particular, the goal is to shore up the case for metacognition against criticism by appealing to the role of extracerebral body in enabling and facilitating these mechanisms and capacities (for this criticism, see e.g. Carruthers 2008, 2009a, Fletcher and Carruthers 2012; Carruthers and Ritchie 2012; Le Pelly 2012, 2014; Langland-Hassan 2014; Carruthers 2017, Carruthers and Williams 2019, Nicholson et al. 2019). How exactly the body is able to fulfil this role, and thus bolster the case for metacognition will be the central topic of subsequent chapters (Chapters 2, 4 and 6). For now, the dominant view of what metacognition consists in will first need to be made sense of.

One distinguishing characteristic of metacognition is a peculiar species of mental entity known as ‘noetic feelings’. Thus, what exactly noetic feelings are and how they work in the overarching cognitive economy must be made sense of before it can be discussed how best to augment metacognition to ward off criticism. This chapter thus serves to introduce and discuss Joelle Proust’s influential theory of metacognition with a specific focus on her account of noetic feelings (1999a, 2003, 2007, 2008, 2009, 2010, 2012, 2013, 2014, 2019).

In order to properly make sense of her theory of metacognition, Proust’s theory of mind will first be discussed by analyzing it as a cybernetic theory of the animal mind (1995, 1999, 2010, 2013). The result will be a thorough accounting of her theory of mind and metacognition, as well as the groundwork needed for making sense of the cognitive architecture underlying noetic feelings that will serve as the foundation for the embodied account of noetic feelings in the following chapters. The goal is thus to mine Proust’s body of work for the strongest possible framework for responding to the various criticisms that have been levied against her theory of metacognition,
while nonetheless remaining true to her proposal, and in so doing, and with the benefit of hindsight, preserve this framework for the argument in later chapters.

In this regard, Proust’s case for construing metacognition as a precursor to self-knowledge will be interpreted as an appeal to the origin of a subject-world distinction in thought, a property distinctive of human cognition (see Section 7.1). That said, it will be discussed how this appeal struggles to address criticisms of the view (see Section 7.2 and Chapter 3; also, see Introduction for a review). Thus, the approach of the present work for shoring up metacognition will be different from Proust’s, though her theory of metacognition will nonetheless function as one of its starting points (the other starting point being the social-scaffolding view; see Introduction). Specifically, her theory will be construed here as pointing the way to an appeal to embodied metacognition, a family of mechanisms and capacities for producing noetic feelings as conscious embodied states of executive control over cognition and behavior, an account which will be expanded upon and defended in subsequent chapters (see Chapters 2, 4, and 6).

1. Introduction

In this chapter, I begin by discussing Proust’s philosophy of mind but mostly as a means to an end, namely that of accounting for her theory of metacognition. Though this analysis will be conducted through a critical prism, it will be the goal of the present chapter to unearth the strongest possible case for her theory. In particular, we will discuss how her account might plausibly resolve two central problems previously introduced, the problem of triviality and the problem of the origin of self-knowledge (see Introduction), problems that will be with us throughout the present work.

This analysis will present her theory as maintaining that metacognition makes possible an origin of a subject-world distinction in thought, which, as will be discussed, has the potential to resolve the problems above, but nonetheless struggles to do so. As will be discussed here and elsewhere (see Section 7.2 and Chapter 3), the problem with this proposal is that it remains nonetheless unclear how exactly the mechanisms of metacognition make this distinction possible in a manner distinct from how other, more cognitively basic mechanisms might also make it possible. What is more, it remains unclear whether such an appeal can help us make sense of the origin of self-knowledge, which is the ultimate explanandum of capacities for metacognition (see Introduction). Thus, though Proust’s approach will be faithfully reconstructed and motivated, it
Chapter 1: What is Evaluative Metacognition?

will not be defended; rather, it will be demonstrated how this approach would likely fall short of fully resolving these problems in the eyes of skeptics.

That said, Proust’s theory of metacognition will be construed as providing the foundation for an approach that might plausibly solve these problems (see Section 8), a case that will developed in subsequent chapters (Chapters 2, 4, and 6). Essentially, this is the case for embodied metacognition: an augmented version of evaluative metacognition that appeals to noetic feelings as strongly embodied, as conscious forms of executive control over cognition and behavior. If the arguments made by the present work are correct, this augmented version has the tools to build the required bulwark to shore up metacognition against criticism, as well as the bridge connecting the phylogenetic to the ontogenetic origin story about self-knowledge (see Introduction).

To introduce Proust’s philosophy of mind, I find it helpful first to introduce her influences and how they have shaped her thought. One notes her influence by early analytical philosophy, in particular Strawson (1959) and Quine (1953, 1956), the former of whom laid the foundation for her distinction between propositional representational systems and feature-involving representational systems (in this regard, see also Glouberman 1976), and the latter of whom paved the way for her thought on different modes of references (de re, i.e. indexing, and de dicto, i.e. representing as; see also Burge 1977).

In this respect, feature-involving representational systems are deployed to make sense of the minds of all animals, human and non-human alike, while propositional representational systems are theorized to be exclusive to (or distinctive of) the human mind. As such, one can understand Proust’s theory of metacognition as a theory about how the former system, with its cognitively modest mode of reference (viz. de re), can serve as a precursor to the latter kind of cognitive system with its cognitively sophisticated mode (viz. de dicto: see Sections 3 and 2, respectively).

Her thought is also influenced by the work of Evans (1982) who formulated conditions that any cognitive system must fulfill in order to possess propositional representational systems and conceptual capacities, while also providing a widely influential account of non-conceptual representations, the latter of which serves as her foundation for thinking about the kinds of mental representations common to all animal minds. Thus, Evan’s work provides Proust with the tools for analyzing the cognitive architecture that enables propositional representational systems, as well as the tools for delineating the human mind from that of other animal minds (see Section 2).

One rather crucial influential source is the work of the metaethicist Gibbard (1986, 1990) who formulated an expressivistic theory of normative discourse, aspects of which readily lend themselves to thinking about the possibility of normativity among non-human animals, one which
forms the foundation for Proust’s thought on the origin of a normativity of an epistemic kind. As we shall see below, the role of metacognition in realizing epistemic normativity is the crux of Proust’s argument for why metacognition ought to be construed as a crucial precursor to more sophisticated capacities for acquiring self-knowledge (see Sections 6 and 7).

Sources outside the realm of philosophical discourse equally influence Proust’s thinking. Her philosophy of mind builds off pioneering work in cybernetics (e.g. Miller et al. (1960) and Conant and Ashby (1970)), as well as Gibson’s (1979) groundbreaking work in ecological psychology. Also, one can readily note the debt her thinking owes to Kahneman and Tversky’s Systems Theory of Cognition (1982). Essentially, Proust’s theory of mind is a cybernetic theory, which appeals to the dual system’s view to distinguish sophisticated cognition (viz. system 2) from a more modest form (viz. system 1). Meanwhile, she employs a modified version of Gibson’s notion of affordances to articulate the structure of system-1-based representations crucial for her account of metacognition and her theory of the prerequisite cognitive architecture for system-2 based mental representations (viz. propositions).

Perhaps most influential for Proust’s philosophy of the mind are contemporary studies in comparative psychology based on the uncertainty monitoring paradigm (see Smith et al. 1995; Smith et al. 2019; see Introduction). These studies are commonly theorized to demonstrate advanced decision-making skills in some non-human animals (e.g. Old-World monkeys and dolphins), skills aimed at mitigating uncertainty and exploiting confidence. To many researchers, these results suggest some species of non-human animal possess a basic rudimentary form of self-knowledge that serves as a phylogenetic precursor to the more advanced forms of self-knowledge in humans, such as those forms of self-knowledge obtained by deploying capacities for metarepresentation and mindreading (for detailed discussion of these studies, see Chapter 5).

Proust’s philosophy of the mind can thus be construed as a philosophy of the animal mind, (both human and non-human animal), as well as a cybernetic theory of the animal mind. Concisely, for her, the mind is a complex network of interwoven control systems, within and across the cognitive hierarchy, in the business of predicting, monitoring, evaluating, and controlling cognition and behavior in varying degrees of specificity. In a slogan: *having a mind is having control.*

As such, her theory of mind is largely compatible with the increasingly influential Bayesian Brain Hypothesis and theory of cognition called ‘predictive processing’ (which also owe a debt to cybernetics), all of which envision the mind as a kind of predictive engine (Knill and Pouget 2004; Hohwy 2013; Clark 2016). As will be discussed in Chapter 5, this overlap presents both challenges and opportunities for metacognition. However, these intersections cannot possibly become the
focus of our discussion until significant progress is made in making sense of metacognition in its own right.

In this chapter, details surrounding Proust’s cybernetic theory of the animal mind will be discussed, in particular how it lays the foundation for her theory of metacognition and noetic feelings. To begin, those aspects of the human mind that distinguish it from the minds of non-human animals will be discussed (Section 2). After this, our discussion will shift to talk about those aspects that the minds of all animals have in common, human and non-human animal (Section 3). In the next section (Section 4), we will discuss the flexibility of the animal mind and this flexibility will be shown to be rooted in the flexibility offered by metamnemonic mechanisms; meanwhile, critical aspects of this flexibility will prompt a discussion of the relevant representational systems.

Those four sections comprise the first part of this chapter (Part 1), together paving the way for the second part (Part 2). Here, details surrounding Proust’s argument for construing metacognition as a significant precursor to self-knowledge will be discussed. Her argument will be shown to depend upon Gibbard’s naturalistic theory on the origin of normativity (Section 5), as well as the cognitive process of calibration that Proust theorizes as giving rise to a normativity of an epistemic kind (Section 6). The chapter thus concludes with a discussion on how Proust’s theory of metacognition is argued to make possible distinctly-human cognition and ends with several points of criticism that will remain pertinent throughout the present work (Section 7). Finally, a brief summary will be provided of those key aspects of Proust’s theory that will serve as the basis for the present work’s case for embodied metacognition (Conclusion).


2. The Human Animal Mind, or the Minimum System Requirements for the Minds of Curie and Shakespeare

Proust’s theory of mind is a theory of the animal mind, in that she believes chief components of the cybernetic architecture pertain to the minds of humans as well as to non-human animals. This is not to say, however, that her philosophy of mind does not distinguish the cognitive capacities of human cognition from those of non-human animals. So, going forward, it will be

---

7 Going forward, when the context does not demand further specification, I shall refer to non-human animals as simply ‘animals’, which is not in any way to deny how we humans are, after all, animals too.
helpful to distinguish first those properties that make human cognition distinct before discussing the properties that human cognition has in common with the minds of non-human animals.

In Proust’s philosophy, a highly sophisticated representational system distinguishes the mind of humans, one that can be described as being endowed with four properties: ‘particularity’, ‘universality’, ‘objectivity’, and ‘generality’. One useful approach to making sense of these properties is to consider them as picking out capacities for producing representations in specific modes: as a particular, as a universal, as about an objective world, with the property of generality referring to the mode by which representations enter into combination with one another to yield complex representations in a highly flexible and open-ended manner.

To refer to these properties individually is not to say, however, that they are necessarily isolatable, neither at the neurobiological, nor at the cognitive, nor at the psychological level of description. Rather, the claim is that these properties reflect distinct explanatory features, individuated in comprehension, that cannot be dispensed with when providing a sufficient account of what makes human cognition distinct.

In order to grasp these properties as enabling sophisticated cognitive feats, underpinned by advanced cognitive mechanisms, it helps to recognize how non-trivial certain basic acts of human cognition are. For example, on my writing desk sits the cup of coffee my brother gave me for Christmas last year. Take note how non-trivial it is that my cognitive system is capable of representing this object as a particular cup, as potentially distinct from all other cups.

Metaphysical questions about whether this cup truly is distinct, and, if so, how to make sense of this distinctness will not concern us (see Thomson 1998); nor will questions in the philosophy of mind concern us about whether the cup is represented as a particular by perception or whether its particularity in cognition arises in virtue of casting a judgment (see Benjamin 2007; Schellenberg 2010). Rather, Proust is interested (and thus so are we) in accounting for both the cognitive capacities and the cognitive mechanisms that make the representation of this coffee cup as an individual object possible, whether this be the result of perceiving or judging it as such.

Next, notice that we are capable of using the term ‘coffee cup’ to refer to a near infinite number of objects irrespective of their individuality, but in respect to some set of universal coffee-cup-like traits that all objects need possess in order to fall under the term. Here, our concern is not

---

8 Strictly speaking, the property ‘objectivity’ picks out how mental representations can be about objects, particular things that persist through time and are the bearers of properties. This manner of construing the property above is not by any means antithetical to how it is construed in this strict sense, however; on the contrary, construing the property as related to an objective world is the essence of the strict construal: a world of objects that are or are not the bearers of specific properties, a matter that is truth-evaluable, objective (see Proust 2013: 116-177).
with certain philosophical questions about whether this cognitive operation occurs in virtue of the deployment of sensorial images, Platonic Ideas, linguistic descriptions, or generative models describing the sensory causes of coffee-cup-like objects (see MacBride 2018). Rather, Proust’s philosophy of mind remains neutral to these proposals and seeks to determine what all presuppose about the cognitive mechanisms and cognitive capacities that enable and facilitate this feat of representing universals (or producing representations in a ‘universal mode’).

Moving on to the next property, notice that if the content of a mental representation is going to be about an objective world, this representation must somehow aim at highly stringent conditions for correctness, namely truth conditions. Here, the goal is not to understand the nature of this relationship (see Shea 2018): do accuracy or truth conditions play a role in determining the computational role of representations? Nor is the goal to make sense of what determines the truth conditions for specific representations (see Schellenberg 2018): is it the capacities deployed, facts about external states of affairs, or externally localized, socio-cultural practices? Rather, Proust’s question concerns the configuration of the cognitive mechanisms and the cognitive capacities for handling representations that aim at truth, however this is thought to unfold or emerge.

Finally, to possess generality a cognitive system must be capable of producing its various representations together in a highly flexible and open-ended manner. For example, in our thinking as humans, the representation evoked by the term ‘coffee cup’ can become easily combined with all manner of other representations, such as the representations evoked by terms ‘desk’, ‘kitchen table’, ‘bookshelf’, ‘the floor’, etc., and still yield a sensible product. Many combinations that do not make sense together (e.g. ‘coffee cup’ and ‘democracy’) can still be entertained in spite of their non-sensical nature (e.g. a mental image of a coffee cup running for President comes to my mind, which would perhaps make a better President than some). Though some combinations do not strike us as sensible no matter how far our imagination might stretch (‘square circle’ and ‘green idea’ serve as good examples of this), emphasis is duly placed on there being a near infinite number of such sensible combinations among representations.

Notice these four properties are vital for explaining the high level of cognitive flexibility enabled by the cognitive capacities of the human mind, one that might be best summed up as the ability to think beyond the confines of the here and now. With particularity, universality, and objectivity one can think beyond the state of affairs present (right here, right now) and think about states of affairs not immediately present in the environment.

Thus, Proust inherits this aspect of her philosophy of mind from Evans (1982), who argued that thoughts about this particular coffee cup as a particular cup are only possible if the subject can
also entertain thoughts about how this cup is distinct from other cups, which are not currently present (see Evan’s well-known ‘generality constraint’; ibid: 99-101). Together with the property of generality, one can construct thoughts of near infinite potential, that is, think about ways the world could be or could have been, also called ‘counterfactual reasoning’, make sense of metaphor and fantasy, determine empirical fact and even eternal truth.

Consider how combining the terms ‘sun’ and ‘Juliet’ in the right way yields thoughts about a Shakespearian drama, while a specific combination of ‘matter’ and ‘energy’ yields thoughts about Einstein’s Special Theory of Relativity. Consider how impossible both the arts and the sciences would be if thoughts were restricted to only the confines of the here and now. Once this (perhaps) pinnacle of cognitive flexibility is reached, so as to require all four of the properties above to explain its feats, the cognitive system becomes capable of producing novel mental representations believed to be distinctive of human cognition, namely concepts and propositions: concepts are those representations that are described by the properties of particularity, universality, and generality, while propositions are those mental representations with concepts as constituents, described by the property of objectivity, which thus aim at truth.

Thus, building upon foundational work in modern analytic philosophy, Proust argues that concept deployment is a cognitive act enabled by representational systems that possesses the four properties above, properties which, ex hypothesi, describe only the cognitive capacities of humans. That said, it is not the purpose of this chapter (or this dissertation) to defend the view that only humans have concepts and propositional thoughts (though, it is a rather common belief: see Clark 2000; Bermúdez 2004; Carruthers 2009). Rather, the goal is to account for Proust’s theory of metacognition and noetic feelings which, in turn, requires providing an overview of her philosophy of mind. In this respect, Proust does not develop a robust account of the cognitive mechanisms that make these four properties of cognition possible, as a full account would require tackling deep philosophical conundrums like those mentioned above, not to mention addressing certain empirical questions, such as the exact nature of their neurobiological underpinnings.

It is clear, however, that Proust envisions these four properties to play an explanatory role in how the representational systems of humans enable the entertaining of truth-functional mental representations. Moreover, and more importantly for present purposes, it is also clear that she believes the family of mechanisms that make metacognition possible also play a significant role in making these four properties possible, and her account of this will be sketched out at the end of this chapter (see Section 7). For now, let us step back and relate our discussion of the four properties of distinctly-human cognition, as well as the levels of cognitive flexibility bestowed by
them, to the topic of metacognition, in particular metarepresentational metacognition (or mindreading).

Metarepresentational metacognition can be thought of as comprising various capacities for representing representations as representations. Often referred to as de dicto reference (in contrast to de re reference, also known as indexing, see Introduction and Chapter 3), this notion of representing as such is reflective of highly sophisticated cognitive capacities, deeply crucial to explaining distinctly-human intelligence and creativity. In order to represent a representation as a representation (e.g. the belief ‘It is raining’ as a belief, which could be either a true or a false about the weather) the mental representation must be represented as aiming at truth, as pertaining to an objective world, as referring to some particular state of affairs that instantiates one or more universals (e.g. RAIN), a state of affairs that could be potentially otherwise (could potentially instantiate some other universal), that is, different from how it currently is, as it were, in the confines of the here and now, so that the representation could potentially be a misrepresentation.

Hence, in Proust’s philosophy of mind, the four properties above make possible metarepresentational metacognition and thus, very crucially for the present purposes, metarepresentational metacognition is the explanandum that evaluative metacognition must seek out if the latter is to be construed as a phylogenetic precursor to self-knowledge. This point will remain throughout the dissertation, and it will be argued that embodied metacognition, an augmented account of evaluative metacognition, ought to be construed as providing the tools for bridging evaluative to metarepresentational metacognition (see Chapters 4 and 6).

By localizing and describing the properties of the distinctly-human cognitive system, Proust offers, what can be described as, ‘the minimum system requirements’ not only of human-level intelligence and creativity but also the minimum system requirements of metarepresentational metacognition, that is, the minimally sufficient properties (from an explanatory standpoint) that any cognitive system must possess in order to be able to produce representations as representations. Now, the goal of the present chapter is to account for Proust’s theory of evaluative metacognition, a family of monitoring and control mechanisms that paves the way for capacities for metarepresentational metacognition. In particular, evaluative metacognition plays a crucial role in producing the four properties that explain the capacity to metarepresent, a sketch of which will be provided at the conclusion of the present chapter, that is, once more essential details are introduced (see Section 8).

Going forward, it is important to recall just how distinct evaluative metacognition is from metarepresentational metacognition (see Introduction). While the former describes nested and
Chapter 1: What is Evaluative Metacognition?

interweaving levels of cognitive control structures, these structures operate below the threshold of conscious awareness and, at most, produce feedback signals that have the potential to motivate and guide conscious control over cognition and behavior. Meanwhile, metarepresentational metacognition involves nested levels of mental content of personal-level mental representations, conscious beliefs about beliefs. So, the challenge for Proust is accounting for how the (conscious or non-conscious) products of cognitive control structures, with their first-order representational structure, explain the origin of conscious second-order mental representations (see Section 8; but also, Chapter 6 for the present approach to meeting this challenge). For now, let us continue introducing Proust’s philosophy of mind by turning to the animal mind, whether it be of a human or of a non-human animal.

3. The Animal Mind, or the Minimum System Requirements for the Minds of Apes, Monkeys, and Dolphins

This section will focus on those features that all animal minds have in common, human and non-human alike. Before it was discussed how human minds deploy concepts that can be combined to form propositions, which are truth-functional mental representations that aim at truth. Intuitively, attributing non-human animals the capacity to entertain the possibility that their thoughts are either true or false would be to overstate the capacities enabled by their cognitive systems. So, not every animal mind has the cognitive system required to acquire and deploy concepts and use them to form propositional representations. But all animal minds, human and non-human alike, are able to produce, and potentially entertain, cognitively less sophisticated forms of mental representations called ‘affordances’.

For our purposes in understanding Proust’s philosophy of mind, affordances describe products of the cognitive systems of all animals. Often, they are deployed by both philosophers and cognitive scientists to explain sophisticated animal behavior without recourse to internal mental representation. Instead of producing a representation, the animal mind reaches out, as it were, and interacts with an external property of the environment, one which the animal’s capacities for action have been honed by both evolution and development to engage (see Gibson 2000). Rather than appealing to complex internal representations of various features of the world to explain an animal’s behavior to e.g. flee the scene, such that there is a mental representation of a predator, a representation of an escape path, etc., this behavior is believed to be sufficiently
Chapter 1: What is Evaluative Metacognition?

accounted for by appealing to the animal’s perception of one or more external properties of the distal environment: e.g. the property of the predator to appear threatening, the property of the escape route to appear inviting. That said, this is different from how Proust deploys the notion of affordance.

For Proust, affordances are not external properties of the world that engage capacities for action; instead, affordances are the representational products of a cognitive system sophisticated enough to possess an animal-level of description (and with it, limited degrees of executive functioning), but not so sophisticated so as to possess the four properties that enable distinctively-human feats of cognition, and so possess metarepresentational metacognition.

Affordances for Proust reflect cognitively light mental representations produced by cognitively light representational systems (and she is not alone in thinking this, of course; see e.g. Wheeler 2005; and for a recent overview, see Dings 2020). As such, affordances do not describe particulars instantiating universal properties, such that their content could be described by the statement, ‘This is the coffee cup my brother gave me’ (which is either true or false of the world). Rather, affordances describe functional relationships between capacities for action (both bodily and mental action) and certain features salient to the organism, either features in the distal environment or features produced by the cognitive system itself (more on these features below).

Thus, Proust envisions two explanatorily distinct cognitive systems capable of producing affordances. 9 A feature-placing system that represents relationships between capacities for action and features in the distal environment, and a feature-based system that represents relationships between capacities for action and features produced by the cognitive system itself. Both of which she distinguishes from the distinctly-human representational system capable of producing propositions. This latter sort of feature-involving representational system, as well as its peculiar species of affordances, ‘cognitive affordances’, will become the focus of our discussion below (Section 5).

Since the representational system producing affordances lacks the four enabling features of distinctly-human cognition, affordances are extremely light on their feet as far as mental representations go. While affordances represent functional relationships between capacities for action and salient features, they do not represent a particular relationship, involving particular capacities or particular features. Nor do affordances represent the relationship as a universal kind; nor are affordances represented as recombinable in a near infinite number of ways with other

---

9 These are labelled as ‘explanatorily distinct’ because of how Proust envisions these two systems as forming a cohesive functional whole in the minds of those who possess them.
affordances. Crucially, despite failing to be truth-functional, they continue to aim at conditions for correctness, just less strict conditions than those that propositions aim at. She does not give these correctness conditions any label, but it is clear that they are epistemically weaker than truth, akin to the accuracy conditions of perception and the satisfaction conditions of desire.

Let’s consider an example. I live in a part of Edinburgh not far from the massive hill, Arthur’s Seat. It is a relative gentle and often forgiving hike to the summit and I often trek to the top to enjoy the beautiful vista of Old Town below and the North Sea out yonder. While hiking along its rocky slopes, I tend to seek out and explore new paths up and down this hill that is so large one would be forgiven for mistaking it a mountain. To translate this into the language of Gibsonian ecological psychology, I tend to seek out and come to perceive new affordances for climbing.

Typically, we would want to say such affordances are not mental representations of the environment but actually external properties of the world that align with capacities I possess for climbing. However, Proust’s notion of affordances is a species of light mental representation that describes a functional relationship between external features of the distal environment (namely, those features that make this slope apt for climbing) and my capacities for climbing (namely, those capacities apt for climbing this slope). Affordances are thus produced in such a manner that key aspects of this relationship are represented at the animal-level of description, so that the creature itself has access to those aspects, in particular those aspects, which are used in multimodal integration and are integral to action selection.

However, it is not as though this representation describes particular features of the environment, features that are individual and distinct from the subject herself, nor does this representation describe capacities as individual and distinct; nor does it describe universal features, nor universal capacities. Rather, the affordance is confined to the immediate here and now of sensory experience, and, as such, does not represent anything other than these features and these capacities (i.e. at this place and at this time) in this functional relationship with the subject’s capacities, thus ruling out the possibility that affordances can be re-combined in a highly flexible and open-ended manner (for a discussion of these indexical links in affordances, see Proust 2015).

In other words, Proust’s affordances are in themselves one-time-use mental representations produced on-the-fly for the immediate demands of the here and now. At their most cognitively sophisticated, affordances serve as indexes of the relationship between features and capacities, pointing the animal’s mind to how these features relate to these capacities (or vice versa), by temporarily representing this relationship in accordance with the momentary demands of the
animal’s active engagement with the environment here and now (see Proust 2015). As such, Proust’s affordances might be more akin to Millikan’s (1995) ‘pushmi-pullyu’ representations than Gibson’s affordances. But describing affordances as one-time-use is a bit disingenuous, since affordances are not divorced from mnemonic mechanisms for storage and retrieval (for more on this, see Sections 4 and 7).

Continuing with the example, it is easy to see how affordances are not truth-functional by redescribing their content in language. Consider the affordance with the content: ‘Climb!’ (with the indexes ‘this’, ‘here’, and ‘now’). This is not a declarative statement, such as “It is raining”, whose content is subject to truth conditions; rather, it is an imperative statement and its content is not truth-functional.10

However, this is not to say affordances lack correctness conditions altogether. For example, if I am too tired or lack the necessary equipment for climbing the slope, my cognitive system might misfire and produce a climbing affordance that misrepresents the functional relationship between the features of the slope and my capacities for climbing, one which motivates me to act but nonetheless dooms me to failure. It is in this sense that affordances are mental representations: affordances are capable of misrepresenting and so aim at conditions for correctness.

Let us therefore take a page from the expressivist Gibbard (1990), whose influence on Proust will be discussed in the next part, and call the correctness conditions for affordances ‘aptness conditions’. Accordingly, affordances are imperatives that issue demands that are either apt or inapt, so that creatures can perform them well or, for whatever reason, not well. This will be revisited below (see Section 6).

Before moving on, let us take stock. Only the human mind possesses a representational system capable of producing propositions (i.e. mental representations aiming at truth). But animals, whether human or dolphin, possess a representational system capable of producing affordances (i.e. mental representations aiming at aptness).

Crucially though, Proust does not deny that the cognitive systems of many non-human animals possesses the sophistication of an animal-level of description, at which point information becomes integrated multimodally for the purpose of selecting context-sensitive action policies, actions that do not immediately succumb to stimulus control and so manifest degrees of executive

10 Of course, it could be true or false that you followed the imperative, but the statement that captures this is not itself an imperative but a declarative statement about the events that followed. Also, while an imperative could be issued that is impossible to comply with, this would not mean we are dealing with a false imperative. Being intertwined with truth conditions requires staking a claim, and imperatives do not claim, they demand. The imperative, ‘Give me cake!’ does not make a claim that you have cake, it just demands you give me some (well, you better find some then!). That said, it is another thing entirely to say an imperative issues an inapt demand (see immediately below).
functioning. So, in denying animals propositional representational systems, Proust does not also thereby deny them access consciousness, so that it stands to reason that non-human animals can become conscious of affordances.

This point will be revisited below (see Section 7 and 8). There, Proust’s theory of evaluative metacognition will be interpreted as pointing to conscious affordances for executive control over cognition and behavior as a key element in shoring up metacognition against the problem of triviality (see Introduction; see Chapter 4 for the proposed solution to this problem and Chapter 5 for an alternative proposal, both of which involve an appeal to consciousness).

In summation, while only humans are capable of thinking about how their beliefs and actions aim at truth, non-human animals, to the extent they are capable of thinking at all, are theorized as, at best, only sensitive to how their belief-like states aim at aptness. Consider the cat that has scurried up a tree, now perched on a limb and softly mewing, to avoid being eaten by the neighborhood dog. To the extent the cat is capable of thinking, the cat thinks that jumping out of the tree is a bad idea. That is, the cat itself has some sense of the aptness of its behaviors, however rudimentary. However, the cat’s capacity to entertain the aptness of its behaviors is bound to the confines of the here and now. That is, after the fact, after its caretaker has coaxed it out of the tree and into the arms of safety, it is not as though the cat can reflect in silent solitude on its behavior, consider whether climbing up the tree had been the right thing to do, whether it could or should have done differently. Clearly, this would be to give the cognitive systems of non-human animals too much sophistication. This level of cognitive flexibility, entertaining hypotheticals and considering counterfactuals, made possible by metarepresentational capacities, is distinctive of human cognition, for better or for worse. But this is not to say that the non-human animal mind is utterly without degrees of flexibility, however modest. In particular, minds equipped with not only feature-placing but also feature-based representational systems (i.e. those systems that produce cognitive affordances) can exhibit degrees of cognitive flexibility that, Proust believes, ought to be considered precursors to the flexibility of the human mind. Let us turn to this now.

4. Flexible Cognition is Rooted in Flexible Metamemory

In the previous section, Proust’s notion of affordances was introduced as cognitively-light mental representations produced by the cognitive systems of the animal mind, which is sophisticated enough to possess a psychological level of description, but not so sophisticated as to
possess the four properties above that enable distinctly-human cognitive feats like metarepresentation. Affordances lack this sophistication as a result of being bound to the immediate here and now of sensory experience.

However, this claim about affordances should not be read as maintaining that the affordance-producing representational systems entail no degrees of cognitive flexibility. On the contrary, Proust envisions sophisticated levels of flexibility to be enabled by these systems, especially the one which produces affordances of a particular kind, cognitive affordances. In fact, improved cognitive flexibility is the central payoff of a cognitive system equipped with mechanisms for metacognition, a point crucial to Proust’s case for construing metacognition as a crucial phylogenetic precursor to capacities for acquiring self-knowledge. So to introduce her theory of metacognition, let us first discuss the source of the cognitive flexibility enabled by metacognition, namely metamemory.

The family of mechanisms underpinning metacognition are chiefly in the business of predicting, monitoring, and evaluating cognitive operations. Together these mechanisms comprise a complex cognitive system capable of performing operations that are described as having three properties that are distinctive of animal cognition or highly intelligent animal cognition (human and non-human animal alike), namely ‘adaptive’, ‘dynamic’, and ‘sensitive to epistemic norms’.  

To make sense of Proust’s theory of metacognition, it helps to compare and contrast these three properties to the four properties above. Whereas the four properties above describe how mental representations are produced (likened to modes of representation), these three properties (adaptive, dynamic, and sensitive to epistemic normativity) describe how cognitive operations are performed, operations as diverse as overt actions (such as climbing a rocky hill) to covert cognitive processes (such as determining the density of a visual stimulus). But what exactly it means for actions and processes to be adaptive, dynamic, and sensitive to epistemic norms, as well as how these three properties are theorized to enable sophisticated degrees of cognitive flexibility which pave the way for the four properties of distinctly-human cognition cannot be adequately discussed until more is said about her theory of metacognition.

---

11 Proust uses ‘adaptive’ somewhat differently than is commonly used in the literature on the philosophy of biology. For her, it does not mean improving biological fitness but rather using information as a causal medium for cognitive regulation (see below).

12 It is not entirely clear from Proust’s discussion whether these properties describe the minds of all animals or just those animals that perform on par with humans in the uncertainty monitoring paradigm (see Introduction). Due to how her aim is often to explain the observed behavior of animals who pass these tests of uncertainty (see Proust 2013: 79-109), I shall interpret her to suggest that these properties in total, especially the third property (sensitive to epistemic norms), describe the minds of those animals who do well in the uncertainty monitoring paradigm, which I paraphrase as “highly intelligent animal cognition” or “distinctly-animal cognition”.
I find it helpful to think of Proust’s theory of metacognition as, first and foremost, a theory about metamemory. Not only did the basis for her theory of metacognition emerge from decades of studies into metamemory (see Hart 1967; Brown 1978; Flavell and Wellman 1977; Nelson and Narens 1990), metacognition, as it exists in its current form, continues to depend upon mnemonic mechanisms in three crucial ways. First, metacognition depends upon access to profiles stored in memory, profiles that describe the prior performance of cognitive operations. These profiles are not unlike performance standards, in virtue of which the monitoring, evaluating, and controlling of cognitive processes are made possible through comparing performance to these internal standards. Thus, without these profiles stored in memory, metacognition would never get off the ground.

The second crucial way metacognition depends on memory concerns the target of metacognition, i.e. what is monitored, evaluated, and controlled, about which predictions are issued. Though these mechanisms are in the business of monitoring, evaluating and controlling overt actions and covert processes, understood in a broad sense to include the domains of perception and decision-making, cognitive operations in these various domains are theorized to occur in virtue of the activation of select mnemonic nodes, so that the mechanisms of metacognition ultimately target patterns of neurocomputational activity unfolding across select mnemonic nodes.

Third, at their most basic the products of these processes of evaluation and prediction are cerebrally-mediated signals that describe the degree to which cognitive operations are (or will be) processed fluently, i.e. processed quickly and/or adequately, relative to descriptions of dynamic activity in corresponding profiles. These signals are often referred to as ‘processing fluency signals’, theorized to be nothing less than patterns of mnemonic activity, or properties of such patterns, (see Paek et al. 2020) that “work as indicators for what normal processing should be for a task in a context” (Proust 2013: 129).

Traditionally, mnemonic cognition is split into two subsystems (see Tulving 1972). On the other hand, there is the operating subsystem for retrieving facts from memory. On the other hand, there is the evaluative subsystem that serves two functions: assessing the resources available for retrieval and, second, evaluating whether the retrieved information matches stored expectations.

Thus, at first blush, Proust’s theory of metacognition is a species of metamnemonic cognition, with the operating subsystem reflecting the so-called ‘object-level’ and the evaluating subsystem corresponding to the ‘metalevel’. However, her theory of metamemory can be easily extended to cognition as a whole, so long as we are dealing not only with memory, but also
perception, reasoning, and decision-making, a point that research into metamemory readily accepts (see Dunlosky and Metcalfe 2008).

Since the mechanisms for metacognition are theorized to monitor and control cognitive operations that span the various cognitive domains, from perception to behavior, this cognitive process is described as a *meta*-cognitive one. It is thus this appeal to the mechanisms of monitoring, controlling and evaluating cognitive operations that transforms Proust’s philosophy of the mind into a theory about metacognition. Though more must be said about whether this appeal is sufficient to qualify this process as metacognitive in a sense crucially related to metarepresentational metacognition, this discussion should suffice for a brief introduction to Proust’s theory (for this controversy, see Introduction, Chapters 3 and 5).

Moving on to the details of her theory that explain noetic feelings, notice that having an evaluative subsystem is highly advantageous from both a computational and a behavioral perspective if the cognitive system is fallible on the whole (which animal cognition presumably is). If uncertainty and error can be monitored and mitigated by the evaluative subsystem, then it can ensure resources are allocated to each task as optimally as possible. Crucially, this process can save massive expenditures of cognitive resources by ensuring the overarching cognitive system is only involved if some aspect of the operation goes awry or if some aspect of it is particularly relevant and needs to be made salient to the subject.

A good example of this process in action comes from metamemory research, the so-called ‘tip-of-the-tongue experience’ (see Hart 1967). A subject will only have this experience, produced in part by the evaluative subsystem, if mnemonic retrieval has failed. More specifically, to the degree that the tip-of-the-tongue experience reflects the production of an error signal, it is only produced in the event of a discrepancy between what is expected to occur and what has actually occurred. In later chapters, our discussion will focus on the degree to which noetic feelings, like those characterizing the tip-of-the-tongue experience, are something more than mere error signals, specifically the degree to which these feelings are strongly embodied in creatures like us (see Chapter 2) and how these embodied episodes lay a plausible foundation for solving the ultimate problem on the origin of self-knowledge (see Chapter 6). But for now, let us continue discussing Proust’s theory of evaluative metacognition and how she believes it makes possible degrees of cognitive flexibility that enable distinctly-human levels of cognitive flexibility.

To understand how this (largely) metamnemonic framework enables the three properties above related to cognitive flexibility, the roots of Proust’s thought in cybernetics will need to be unearthed, in particular the basic building blocks laid by the seminal work of Miller et al. (1960),
which famously pushed the field of psychology beyond behaviorism toward cognitive psychology. Building upon previous theories that describe reflex arcs at the juncture between stimulus and behavior, Miller and colleagues proposed replacing reflex arcs with sophisticated monitoring devices that control this relationship and yield more flexible behaviors, feedback loops referred to as ‘Test-Operate-Test-Exit units’ or TOTE units.

Proust describes the operating subsystem as composed of various TOTE units with two-way information flows. One flow carries commands that describe cognitive operations. It travels from the top of the cognitive hierarchy, where mnemonic performance profiles are stored, down to the bottom of the hierarchy, where commands meet effectors. The second flow carries feedback signals that travel in the opposite direction (i.e. from the bottom of the hierarchy back up to the top). These feedback signals, called ‘reafferences’, are produced during both the performance of overt actions (e.g. the action of climbing a steep cliff), as well as during the performance of covert processes (e.g. the process of determining visual density), carrying with them information about how well cognitive operations were performed.

Reafferent signals serve as input to the evaluative subsystem which, in turn, compares the information described by them to information stored in the profiles, effectively producing another layer of feedback. This second layer of feedback is often referred to as an ‘error signal’, or as mentioned earlier, a processing fluency signal. In the event of a discrepancy between observed reafferences and stored expectations, the evaluative subsystem thus produces a processing fluency signal that indicates either worse-than or better-than expected performance.

How exactly mechanisms for metacognition make possible one of the three properties of distinctly-animal cognition above can now be clarified. Proust categorizes cognitive operations as adaptive so long as information is the causal medium for the overarching process of cognitive regulation. To see the import of this for flexible regulation, consider first a much more basic control mechanism that does not use information as a causal medium.

For example, a simple thermostat operates by a bi-metallic strip that bends in a particular direction depending on the temperature of the environment. The obtaining of this physical fact is what enables the strip to function as a control mechanism of e.g. an air conditioning system and thus it acts as a monitor and a regulator of the local temperature. Meanwhile, the physical interaction between the strip and the local temperature exhausts whatever informational structure might exist between them. The regulation of the local temperature is thus explained in virtue of the physical facts that obtain between the control unit and the thing it regulates, in such a manner that the informational structure, itself, plays no explanatory role in this regulatory process.
Regarding TOTE units, their mere physical configuration cannot suffice to explain their functional role as control mechanisms. One needs to appeal to stored profiles describing the expectations of cognitive operations, the discrepancy between observed reafference and its expectation, and the production of two layers of feedback signals to explain TOTE units as control mechanisms. Having information as a causal medium for cognitive regulation means every stage of this regulatory process can be revised and updated to account for change (both in the distal environment and in the cognitive landscape). That is, systems of cognitive control are adaptive, changing in order to improve the performance of cognitive operations (that is, if all goes well) in light of feedback. Notice how this is missing from the bi-metallic strip due to how its function as a control mechanism is a consequence of the obtaining of facts governed by physical laws, which are, by their very nature, laws and so do not give rise to adaptations.

The conclusion is thus that TOTE units adapt due to the deployment of internal models that describe specific input-output couplings, i.e. stimulus-behavior couplings, selected for in varying contexts in virtue of feedback signals that describe the adequacy of these models. Selective pressures ensure that these couplings are in competition with one another over providing the best description of their target operation.

Such competition among candidate models explains the second property of distinctly-animal cognition, dynamic. Cognitive operations are said to be dynamic so long as this competitive process between vying models of stimulus-behavior-couplings enables the cognitive system of animal minds to continually learn from its engagement with the environment, enabling the system to develop ever-changing (viz. dynamic) profiles of reafferent information sensitive to varying degrees of success and failure, tailored to specific courses of action. Thus, metacognition arises from mnemonic mechanisms for the storage and retrieval of observed reafferences for various types of tasks, handling information that is constructed, retained, and retrieved in internal models of the events to be controlled (see Proust 2013: 16).

Based on research into motor control, Proust categorizes internal models into two types. On the one hand, there are forward models that describe the causal relationship between commands and their sensory consequences, and so enable the prediction of reafferent feedback. On the other hand, there are inverse models that describe the transformation of sensory consequences back into commands, and so enable the selection of context-sensitive operations.

This framework for motor control thus serves as the basis for metacognitive control. But the difference lies in how internal models for the metacognitive variant describe overt cognitive acts (e.g. acts of perception, memory, reasoning, and decision-making) and their underlying covert
processes (e.g. the determination of the density of a visual stimulus) and any consequences these actions and processes might have for the internal cognitive landscape (e.g. changes in patterns of neurocomputational activity).

That said, metacognitive control is only reflective of a more sophisticated species of control that is central to Proust’s theory of evaluative metacognition, namely ‘metacognitive governance’. This more robust species of metacognitive control enabled by mechanisms for metacognition is tied to the third and final property of distinctly animal cognition above, namely sensitive to epistemic norms. As a reminder, this final property is theorized to belong to only the most intelligent of animal minds, specifically (but not exclusively, e.g. those who have not been so tested) those animals whose performance is on par with that of humans in the uncertainty monitoring paradigm (see Introduction).

Proust tells us, “The input of metacognitive governance consists in appropriate feedback i.e. sufficiently informational and well-calibrated feedback, and its output consists in a motivation, based on the intensity and dynamics of its associated feeling” (ibid: 21; emphasis added). This thus raises questions about what exactly it means to say that feedback is well-calibrated, a topic that will concern us in the next section below (Section 5). But before details surrounding metacognitive governance and its relationship to the third property of distinctly-animal cognition can be thoroughly discussed, it will be helpful to summarize what has been discussed thus far.

We began by asking about the details surrounding Proust’s cybernetic theory of mind, essentially a system of interweaving layers of cognitive control. Specifically, we asked how such systems enable context-sensitive behavior through adaptive and dynamic control. Adaptive control results from the configuration of feedback loops that enable the cognitive system to learn from its environment, from its successes and failures in coping with it, effectively giving rise to a dynamic process of creating and re-creating models of stimulus-behavior couplings, which, in turn, enables the selection of context-appropriate actions and processes (dynamic inverse models), as well as the evaluation and prediction of actions and processes (dynamic forward models).

Adaptative and dynamic cognitive operations reflect an essential part of the story of how the cognitive systems of animals can begin to approach the level of flexibility of human intelligence, but it still unclear how metacognition can pave the way for the four properties of distinctly-human cognition. So, the second part below will need to explain the remaining, third property of being sensitive to epistemic norms, which goes hand-in-hand with metacognitive governance and the calibration process. Meanwhile, it needs to be clarified how all this relates back to our previous discussion of cognitive affordances and noetic feelings. The details surrounding epistemic
normativity and metacognitive governance will be discussed in the next section (Section 6) and the calibration process in the one after that (Section 7), while cognitive affordances and noetic feelings will only reenter our story last (Section 8).

Part 2: Proust’s Case for Construing Metacognition as a Precursor to Self-Knowledge

5. Metacognitive Governance and the Origin of a Normativity of an Epistemic Kind

In this section, Proust’s notion of metacognitive governance is analyzed in detail, in particular it is addressed how metacognitive governance relates to the property that cognitive systems equipped with mechanisms for metacognition possess, namely sensitivity to epistemic norms. In so doing, it must be explained how metacognitive governance is distinct from mere metacognitive control and how its distinguishing features account for noetic feelings as distinct from mere processing fluency signals. This explanation thus sets the stage for Proust’s appeal to metacognitive governance and noetic feeling as justifying metacognition as a precursor to capacities for self-knowledge. Our discussion begins below with how metacognitive governance enables a species of normativity of an epistemic kind, the critical feature of metacognitive governance that distinguishes it from mere metacognitive control. Ultimately, this analysis converges on the process of calibration, which accounts for how noetic feeling emerge from processing fluency signals.

Epistemic normativity is for Proust all that is crucially involved in both the acquisition, fine-tuning, and deployment of capacities aimed at regulating both covert processing and overt actions according to norms of an epistemic kind, examples of which are ‘accuracy’, ‘exhaustiveness’, ‘fluency’, and ‘truth’. In thinking about epistemic normativity, it might be natural to think that such sophisticated forms of rule-governed behavior are distinctive of humans. For only those subjects embedded in socio-cultural environments, in which behavior is reinforced according to epistemic norms, are subjects whose thinking can be sensitive to epistemic norms, and since the only subjects who have such rich socio-cultural environments are humans, only the minds of humans are sensitive to epistemic norms (see Heyes et al. 2020). This line of reasoning falls out of the influential social-scaffolding view introduced above (see Introduction).
That said, notice this conception of what epistemic normativity consists in suggests its source components are in subjects’ interactions with their socio-cultural environment, rendering the role of the cognitive system not proprietor, but recipient and benefactor of epistemic normativity—a kind of conduit, or clay to be molded. Proust interrupts this line of reasoning by arguing that some crucial source components of epistemic normativity originate from within an individual’s cognitive control structures. While it may be true that self-knowledge has its origin in the socio-cultural practices that forge a collective epistemic agency (a view that the present work assumes is correct), it might nonetheless the case that specific features of cognitive systems are part of the explanation for how these rich socio-cultural practices emerge (see Introduction). Thus, Proust’s theory of metacognition paves the way for thinking about how the minds of some non-human animals might be regulated by norms of an epistemic kind.

As already introduced, Proust’s philosophy of mind is heavily influenced by research in comparative psychology, in particular research based on the uncertainty monitoring paradigm (Smith et al. 1995; Smith et al. 2019). This paradigm has succeeded in demonstrating advanced decision-making skills in some non-human animals, skills aimed at mitigating uncertainty and exploiting confidence. Recall from above (see Introduction) that the paradigm consists in human and non-human animals confronted with making strategic use of an opt-out mechanism to avoid discriminating target stimuli in trials designed to be extremely hard to succeed in.

Researchers believe that subjects who succeed in these tasks demonstrate an awareness of how confident or uncertain they are and thus demonstrate a rudimentary form of self-knowledge that could serve as a precursor to the kind of self-knowledge obtained by deploying mindreading capacities distinctive of human cognition. Any sufficient explanation of such capacities, however ultimately spelled out, Proust argues, ought to appeal to a sensitivity to epistemic norms, in particular the epistemic norm of fluency, which qualifies as an epistemic norm due to how it approximates truth (more on this below).

Though what exactly ought to be made of these advanced decision-making skills is deeply controversial, with some detractors arguing this enhanced decision-making performance can be reduced to mere associative learning (see Le Pelley 2012), it is often claimed that the uncertainty monitoring paradigm demonstrates that some non-human animals possess a rudimentary sense of knowing that they know (see Smith et al. 2019). However, take note, as metacognition has been described thus far, it remains unclear how the mechanisms of metacognition enable and facilitate

---

13 To refer back to the analogy above, the claim is that certain features of cognitive systems are analogous to the chicken-like creature that produced the egg (the social-cultural practices) that produced the chicken (self-knowledge), namely those features, according to Proust, which enable metacognitive governance.
anything beyond basic feats of cognition, such as reinforcement learning. As such, it is anything but clear how such mechanisms are crucial precursors to metarepresentational metacognition, mindreading, and self-knowledge (for more on this exact point, see Chapter 3). But what is clear is that Proust believes the status of metacognition as a phylogenetic precursor to self-knowledge depends upon metacognitive governance (i.e. a sophisticated form of metacognitive control), which ought to be construed as implementing a rudimentary form of epistemic normativity governing the transactions of the cognitive system, thereby transforming processing fluency signals into their more cognitively robust cousins, noetic feelings. As such, the crux of Proust’s argument hinges on what exactly metacognitive governance consists in. So, let us discuss this now.

Proust borrows and repurposes the idea of metacognitive governance from the prominent metaethicist Allan Gibbard (1990) who introduces the concept of ‘normative governance’ to expand upon and clarify his naturalistic theory on the origin of normativity. Gibbard’s theory begins with the premise that normative judgments arose from our need as a species to coordinate and cooperate with one another. If true, this suggests that rudimentary aspects of our as-of-now sophisticated capacities for dealing with norms is shared by some non-human animals, especially those whose survival is dependent on predicting conspecific behavior.

According to Gibbard, normative governance can manifest itself in two distinct ways: either an animal can be said to be ‘in the grip of a norm’ or the animal can be said to exhibit ‘acceptance of the norm’. Whereas Gibbard believes any capacity for accepting norms is dependent on language and so peculiar to humans, the former manifestation of normative governance, being in the grip of norms, is shared between humans and non-human animals alike. Thus, a crucial starting point for Proust is Gibbard’s theory that these two manifestations of normative governance are made possible by distinct cognitive capacities unpinned by distinct cognitive mechanisms, the first implemented in all animal minds and the second implemented only in the minds of humans.

To illustrate how being in the grip of norms is distinct from, and more cognitively basic than, accepting norms, Gibbard discusses the famous experiment on obedience conducted by Milgram (1974). In this seminal experiment, subjects are instructed to administer electric shocks to another subject while being told these shocks would become increasingly painful, even deadly. Secretly working with the experimenter, the supposed victim is never actually shocked, however. Somewhat surprisingly, two-thirds of subjects do all they are instructed to do, even though they increasingly become upset and protest against doing it.

Gibbard interprets these results as exposing an internal conflict between two norms: the norm of non-harm and the norm of cooperating. Speaking about this kind of normative conflict, Gibbard
writes, “[o]ften what we experience is a conflict between our “better judgment” and powerful social motivation” (1990: 58). For Gibbard, the fact that subjects actively protest their compliance with the experimenter’s instructions, on the one hand, while nonetheless deciding to comply with those instructions, on the other, reveals a deep insight into both human and animal psychology, exposing the phylogenetic origin of our normative systems.

While subjects do not accept the norm of cooperating, as they explicitly state they believe that it is wrong to comply with the experimenter’s instructions, they also do not take it to override or cancel out the norm they do accept, that of non-harm. This means, not only does Milgram’s experiment reveal that subjects can be simultaneously sensitive to two distinct norms (both the norm of cooperating and the norm of non-harm), but it also exposes that accepted norms are not always the ones motivating and guiding behavior, and this can be true even when the norms that are actually driving behavior are explicitly rejected. In other words, while we might accept and endorse one norm, we might, nonetheless, be motivated and lead by another norm, one which is in direct conflict with the norm we do endorse.\footnote{One might question whether this is a plausible reading of what is demonstrated by Milgram’s experiment, suspecting that subjects would rather sheepishly comply with the experimenter’s instructions rather than straightforwardly disown or reject the norm of cooperating. But this is not what Milgram’s experiment showed. Famously, as the shocks seemed to get worse, subjects would refuse to go on, but if the experimenter said that they (the experimenter) would assume responsibility, two-thirds of subjects would decide to continue on administering the shocks (see Miller et al. 1995). This suggests, as Gibbard construes it, that subjects endorse the norm of no harm while complying with the norm of cooperating, indeed even though it directly conflicts with the norm endorsed.}

Thus, as Gibbard understands it, Milgram’s famous experiment reveals that we can be guided by a more primitive but nonetheless normative faculty, one that has the potential to exert a strong influence over behavior, effectively holding it in its grip. He proposes a psychological hypothesis for explaining the discrepancy in motivation exposed by Milgram’s experiment, one that is based upon two distinct motivational systems: the normative control system, which is “linguistically infused… that evolved because of the advantages of coordination and planning through language” and the animal control system, which is shared with some non-human animals and has the potential to motivate, demotivate, shape, and constrain behavior.

However, referring to the distinctly-human motivational system as the normative one is somewhat of a misnomer, as Gibbard views animals as capable of norm-conforming behavior. But for specific reasons, he does not think animals can be motivated by norms; for this would require that animal minds represent norms as such (i.e. de dicto reference and metarepresentational capacities; see Introduction), which extends beyond their cognitive abilities.
To illustrate both the differences and the commonalities among humans and non-human animals, Gibbard introduces the elaborate ritual that ensues whenever two dogs meet on neutral ground. Their interaction follows regular patterns that Gibbard believes exhibit a form of rationale that emerged through the pressure of natural selection for coordinating behavior among members of the species. “[I]n this special sense the beasts “follow rules” for social interaction. They have not, of course, decided to conduct themselves by these rules...” (ibid: 69). For deciding to conduct oneself according to a rule as a rule requires, “the capacity to accept norms[, which] is peculiarly human and depends on language” (ibidem).

Regarding Gibbard’s hypothesis, the normative control system, as a cognitive faculty, depends on language because it depends on propositional thought that produces representations as representations, specifically those describing norms as norms. This capacity thus has the effect of enriching a creature’s decision-making repertoire and allowing motivation based on norms as such. Thus human behavior comes to be guided by norms in a manner distinct from how the behavior of non-human animals can be said to be guided by norms.

Whereas animals can only go with the normative flow, as it were, humans can opt to go with the flow or become an eddy against it. We can deliberate and decide which norms to conform behavior to. We are not at the mercy of the here and now; instead, we are capable of decoupling our thoughts from the immediate and exercising control over what follows. This would be impossible, Gibbard suggests, if humans were not endowed with a capacity for representing norms as such, a capacity, we should observe, is enabled by metarepresentational metacognition.

However, this talk about what being in the grip of norms does not consist in can only take us so far in coming to terms with what it does consist in. We thus need a positive account of what being in the grip of norms actually is and how e.g. it is different from the behavior of planets and how it can said that their behavior follows the laws of physics. Here, Gibbard describes being in the grip of norms as a process of internalizing norms constituted as much by behavior as by emotion: “a norm prescribes a pattern of behavior, and to internalize a norm is to have a motivational tendency of a particular kind to act on that pattern” (1990: 70). For Gibbard, tendencies are evolutionary adoptions for coordination, while emotion (or affect more generally) explains the motivation to perform the coordinated act: “When a person’s emotions tend to follow a pattern in this way, we can say that the person internalizes the norm that prescribes the pattern” (ibid: 71).

In other words, Gibbard provides a framework for making sense of what has been called ‘tacit knowledge’, i.e. knowledge that cannot be explicitly expressed, but is nonetheless essential for the successful performance of action (see Stich 1978). On the one hand, evolutionary pressures
ensure that behavioral patterns are tacitly learned by the subject, while, on the other hand, emotions ensure subjects are motivated to engage those patterns. It is now clear how being in the grip of norms is different from planetary motion obeying laws: only subjects can be said to learn, however tacitly, the rules prescribed by norms through undergoing emotions that motivate compliance.

Thus, being in the grip of norms amounts to being motivated to engage in behavioral patterns with specific rules prescribed by norms, behaviors which emerge either phylogenetically through natural selection or developmentally through enculturation. Crucially, norms of the animal system lend themselves to be construed as imperatives to act, analogous to how Proustian affordances are construed above (see Section 3). Thus, the animal control system, which motivates norm compliance and drives behavior, is in many ways functionally congruent to the representational system that produces Proust’s brand of affordances.

Though Proust does not explicitly state this, her cybernetic theory of the animal mind can be interpreted as accounting for the cognitive mechanism underpinning Gibbard’s animal control system. It is namely the combination of both feature-placing and feature-based representational systems that account for capacities that enable being in the grip of norms. Specifically, the feature-placing system accounts for being in the grip of instrumental norms (also called ‘norms of utility’ and ‘pragmatic norms’), while it is the feature-based representational system that accounts for being in the grip of norms of an epistemic kind.

For example, whenever a cognitive system represents an affordance, the subject is met with an imperative to act, and, as such, the subject can be said to be in the grip of whichever norm prescribes the rules that regulate the motivated action. For example, the affordance with the content ‘Climb!’ (this, here, now) can be understood as called forth by the cognitive system during its processing of rules about how best to climb the relevant environmental features, rules that are prescribed by the instrumental norm for climbing such surfaces.

While instrumental norms prescribe rules for governing those behaviors that aim to change the world because some change is desired (see Proust 2013: 6, 27, 152), epistemic norms, on the other hand, are “standards of optimal information acquisition and transfer in a cognitive system” (ibid: 312). As such, epistemic norms prescribe rules for governing those behaviors that aim to obtain information to aid or augment cognitive processes (see the distinction between pragmatic and epistemic actions discussed in Clark and Chalmers 1998).

Therefore, in order to grasp how norms of an epistemic kind become internalized, we must thus turn to feature-based representational systems. As already introduced above (see Section 3), these systems are distinct in that they represent relationships between actions and features _internal_
to the cognitive system itself, thus producing affordances of a cognitive kind, with contents such as ‘Discriminate!’ (this, here, now). Notice how the act of discrimination is a mental act (e.g. the act of discriminating the larger of two visual stimuli), one which is governed by rules about internal features to the cognitive system, e.g. rules for determining the internal threshold at which point external stimuli become discriminable. As a mental act, it obviously does not have the aim to change the world but is rather more akin to what Clark and Chalmers (1998) call an ‘epistemic act’ as it aims to aid downstream cognitive processing, such as those which support recognition and search.

The task is now to determine how feature-based systems arise and how they produce cognitive affordances that place subjects in the grip of epistemic norms. For, Proust believes, this is what ensures mechanisms for metacognition are more cognitively robust than basic cognitive mechanisms, such as those which underpin capacities for (mere) reinforcement learning. As such, the appeal to epistemic normativity comprises Proust’s proposal for how best to solve the problem of triviality discussed earlier (see Introduction). But before this proposal can be thoroughly discussed (see Section 7.1 and 7.2), the cognitive calibration process must be introduced. This is because it explains how norms of an epistemic kind are able to regulate behavior.

6. The Cognitive Calibration Process, or How Subjects Become in the Grip of Epistemic Norms

As introduced in the previous section (Section 5), Proust seeks to account for the cognitive capacities of animals that are able to take advantage of the opt-out mechanism in the uncertainty monitoring paradigm, studies which, she believes, provide evidence of metacognition in non-human animals. If success in this paradigm is to be accounted for, Proust maintains, two questions will need to be addressed, whose answers will illustrate two essential features of cognitive systems capable of employing norms of an epistemic kind and thus enabling success in this paradigm.

First, Proust asks about the minimum system requirements for the capacity to determine whether one stimulus is discriminable from another. For without this capacity, subjects could never succeed in choosing which stimulus is the larger, denser, etc. Second, she asks about the minimum system requirements for the capacity to form reliable predictions about the system’s own functioning. For without this capacity, subjects could never strategically opt-out of hard-to-discriminate trials. Indeed, these two capacities explain the cognitive calibration process.
Recall that in the uncertainty monitoring paradigm (see Introduction), both human and non-human subjects are often presented with two distinct stimuli and tasked with discriminating which of the two is the larger etc. In the first phase of the experiment, subjects are forced to choose by pressing one of two buttons (or levers or a spots on a touch-screen monitor) corresponding to either stimulus, a task which becomes increasing difficult, if not impossible, once the stimuli begin to resemble each other. In these hard-to-discriminate trials, performance plummets.

Also recall that in the final phase of the experiment, subjects are presented with an additional, third option to choose from, namely the opt-out response, which allows subjects to skip the current trial. Clearly, it is in their best interest to skip those trials for which discrimination is most difficult, and this strategic use of the opt-out response is only observed in some non-human animals (for a recent review, see Smith et al. 2019). Thus Proust argues, success in the uncertainty monitoring paradigm amounts to having the cognitive capacity to determine those trials for which discriminatory capacities will succeed compared to those trials for which discriminatory capacities will fail, as well as the cognitive capacity to act on this determination.

As stated above, Proust is interested in accounting for the cognitive system that explains this feat and she begins by asking about the minimum system requirements that would enable the capacity to determine whether one can discriminate one stimulus from another, for example, in terms of comparative visual length. For this, her answer is the cognitive system must be sensitive to having comparatively more evidence for one stimulus over the other, made possible by having differential feedback. In order words, the cognitive system must have feedback from the first stimulus, as well the second stimulus, and then be able to compare these two feedback signals.

Second, Proust asks about the minimum system requirements for a capacity to form reliable predictions about the system’s own functioning. The mechanism that enables this capacity, Proust believes, is the final key to unlocking the specifications of the cognitive system that enables success in the uncertainty monitoring paradigm. This is because, in order for a subject to make strategic use of the opt-out response, the subject must possess a cognitive system that can form predictions about its own discriminatory capacities, that is, cast predictions about its own cognitive operations, while making essential features of those predictions salient to the subject in order to motivate and drive her behavior of choosing the opt-out response. For this, the cognitive system must possess internal models describing performance profiles of cognitive operations, like those that form the basis of perceptual discrimination, and be able to represent an imperative to act so as to behave in accordance with the information described by those profiles.
It is not enough to explain success in uncertainty monitoring paradigm by appealing to a cognitive system that compares two feedback signals about the first and the second stimulus. In order for the act of perceptual discrimination to be reliable, the cognitive system must also be capable of comparing the on-going performance of cognitive operations to internal standards that reflect the norm of those operations: “a system cannot form reliable predictions about its own cognitive functioning without having antecedently stored feedback from many trials in the same task” (2013: 21). This norm is an epistemic norm (and not merely an instrumental one), since it prescribes rules that govern behavior that aims to acquire information to facilitate cognitive processing.

Well-calibrated feedback is thus about the employment of evaluations of cognitive operations which reflect the norms of operations, while representing evaluations as imperatives to act in accordance with prescribed norms. Accurate and reliable evaluation of cognitive operations (i.e. evaluative metacognition) thus ultimately depends upon capacities for comparing current cognitive performance to internal standards that describe norms of cognitive performance.

In other words, the strategic use of the opt-out response is made possible by a cognitive system that possesses performance profiles that describe when perceptual discrimination is normally possible. This is not a mere convenient choice of words, but rather evidence of the congruence mentioned above: Proust’s philosophy of mind spells out the cognitive commitments of Gibbard’s theory on the origin of normativity. Proust’s cybernetic theory of the animal mind effectively describes the process by which norms are internalized and her theory of metacognition describes the process of internalizing norms of an epistemic kind.

Let us now consider how this evaluative and motivational process unfolds within the uncertainty monitoring paradigm. When confronted with two stimuli of in-discriminable density, the feature-based representational system will produce a cognitive affordance for opting out, one that has the content ‘Don’t discriminate!’ (with the indexes ‘this’, ‘here’, and ‘now’, and an index to the required action program). Thus, after training, subjects learn to associate this imperative with the opt-out response (of course, some level of associative learning is at play here; for a discussion of the evidence that suggests this is not mere associative learning, see Smith et al. 2019).

Notice that the cognitive affordance is downstream from several layers of prerequisite processing. There is the initial layer of reafferent feedback generated by comparing ongoing operations to internal performance standards. Recall this initial feedback is used to generate another layer of feedback, namely a processing fluency signal that is, in turn, generated by discrepancies between this initial feedback and stored expectations. In other words, the cognitive system has an
internal model for monitoring and controlling cognitive operations, as well as an internal model for determining how current operations compare to internal standards across similar operations.

In the event an affordance for opting-out is produced, the cognitive operation related to perceptual discrimination failed to meet expectations. This means, the processing fluency signal is one of low fluency that, once used for the purpose of selecting action policies, comes to represent the inaptness of the act of discrimination, thereby traversing the psychological threshold and transforming into a cognitive affordance with the content “Don’t discriminate!”.

At this level of description, Proust theorizes, processing fluency is integrated in a multimodal manner, so that these signals come to recruit unconscious and/or conscious patterns of bodily arousal (see somatic markers; Damasio 1994; Damasio et al. 1996). As I interpret it, the upshot of this process is the transformation of the signal into a noetic feeling of uncertainty (or its inverse, confidence), accompanied by bodily arousal and valence, motivating and guiding the animal’s opt-out response, and the physiological process by which cerebrally-generated processing fluency signals come to recruit patterns of bodily arousal will be the central theme of this dissertation. The present work will ultimately argue that some noetic feelings are made conscious through an extracerebral format (see Chapter 2) and an appeal to these strongly embodied noetic feelings is able to solve the problem of triviality facing metacognition (see Chapter 4), as well as provide the tools for making inroads into solving the ultimate problem on the origin of self-knowledge (see Chapter 6).

But for Proust, what is crucial about this process is how the operative norm is not merely instrumental but an epistemic norm. By supporting the monitoring and controlling of the feature-based representational system that produces cognitive affordances, mechanisms for metacognitive governance make possible functional relationships between cognitive actions and salient features produced by the cognitive system itself. For example, the affordance with the content “Don’t discriminate!” makes salient the relationship between capacities for perceptual discrimination and those internal features that enable discrimination, namely the neurocomputational features of differential feedback (e.g. having comparatively more evidence for one stimulus over the other).

Thus, metacognitive governance yields representations that not only motivate the animal to avoid punishment and obtain rewards, but also motivate the performance of overt cognitive acts, such as acts of perception, discrimination, and identification, i.e. acts governed by rules prescribed by epistemic norms. In short, Proust proposes that the mechanisms of metacognition as a whole explain the origin of capacities for forming beliefs about the self versus beliefs about the world. For example, the belief-like state that a subject is in when presented with a cognitive
affordance for discriminating can be construed as a belief-like state about the subject’s own capacities. Thus, in some basic sense, the complex psychological state characterized by a cognitive affordance is about the subject herself (i.e. *de re* rather than *de dicto*). This crucial aspect about metacognition will be revisited later on (see Chapter 4). There I shall argue for construing this belief-like state as a proto self-belief state that serves as a crucial precursor to states of full-blooded self-beliefs.

Of course, it would be a stretch to claim that this subject is in a position to be motivated by an epistemic norm *as such* (see *de dicto* reference, metarepresentational metacognition, and mindreading discussed in the Introduction). Clearly, this would outstrip the cognitive capabilities of non-human animals that perform well in the uncertainty monitoring paradigm. But this is not what Proust wishes to claim. Rather, her claim is far more subtle: the plausibility of a cognitive system capable of presenting subjects with cognitive affordances that motivate and guide behavior to conform to norms of an epistemic kind, she believes, ought to be sufficient reason to consider such systems as significant precursors to metarepresentational systems that enables the explicit representation of self-beliefs.

Consequently, the present work treats Proust’s case for construing metacognition as a crucial phylogenetic ingredient to capacities for self-knowledge as the initial starting point, motivating the present case for embodied metacognition. The view offered by the present work is thus an augmentation of Proust’s account of evaluative metacognition, one which aims to meet various criticisms levied against it (see Chapters 4 and 6). For now, let us motivate what I take to be the strongest possible case for Proust’s theory of metacognition (see Section 7.1), as well as discuss its most challenging criticisms (see Section 7.2). Thereafter, we shall discuss those key features of Proust’s theory that form the foundation of the present work (see Section 8).

### 7. Motivation and Criticism of Proust’s Theory of Evaluative Metacognition

Our discussion has come full circle, back to questions raised as the start about the nature of the cognitive system that could make possible the four properties of distinctly-human cognition (see Section 2). Proust believes such highly flexible cognitive activity is enabled by adaptive and dynamic models describing the performance of cognitive acts, models which prescribe epistemic norms. Although such flexibility is a far cry from what is offered by metarepresentational capacities,
Proust believes, nonetheless, that this flexibility ought to be construed as a crucial precursor to the flexibility offered by distinctly-human cognition. So, given the analysis and interpretation of her philosophy of mind discussed above, let us make the strongest possible case for this view.

7.1 Motivation

Begin by noticing that mental representations are not generally tokened solely for their own sake but indeed arise as a response to some real-life state of affairs. Consider entertaining the propositional representation with the content ‘I shall walk’ in a specific context, such as deliberating about whether to walk to the local supermarket during a light rain shower or to take the bus or to wait it out. Notice that, not only does it make sense to say that the circumstances of the mind-independent world present a bodily affordance for e.g. walking, but it also makes sense to say that the mental circumstances, mapped out by deliberating about whether to walk, or to take the bus, or to wait it out, present a cognitive affordance that motivates the act of judging to walk, rather than to take the bus, or to wait it out.

With this example in mind, let us thus draft up a sketch of how feature-based representational systems, described by Proust’s theory of metacognition, are related to propositional representational systems, distinctive of human cognition. In this respect, notice that the cognitive affordance above represents the act of judging and not explicitly the content of the judgement itself: the affordance is simply ‘Judge!’ along with its appropriate indexes, ‘this’, ‘here’, ‘now’, and an index for the action program (see Proust 2015). This means, the content of the judgement, produced by the propositional representational system, is linked to the affordance via the index ‘this’ which points the subject to the content of the proposition (i.e. it represents the content in the de re mode of reference). Though it is now clear from this illustration how the two representational systems complement each other, it is nonetheless difficult to see how the flexibility of the feature-involving system enables the flexibility of the proposition-involving system, as I interpret Proust as suggesting.

To better grasp this crucial component, let us address the chief question about whether mechanisms underpinning metacognition play a crucial role in producing any of the four properties distinctive of human cognition. In sketching an answer, let us take the act of judging as our candidate cognitive action. Let us also assume the feature-based system, which produces cognitive affordances, routinely produces and stores affordances for judging various discriminanda (e.g.
visual discriminanda of various features: shapes, colors, etc.; and auditory discriminanda of various features: volume, frequency, pitch; etc.). The question then becomes whether having such a store of adaptive, dynamic, and epistemically sensitive models describing various cognitive affordances enables one or more of the four properties distinctive of human cognition: particularity, universality, objectivity, or generality.

As I interpret Proust’s theory, at the heart of metacognition pulses an epistemic normativity that could serve as a crucial ingredient for structuring animal cognition so as to develop the property of objectivity, in particular by laying the foundation for a subject-world distinction in thought. This is ultimately due to how processing fluency signals are at the core of metacognition, signals which, as Proust argues, aim at accuracy and even approximate truth. An advancement in regulating cognition and behavior according to the epistemic norm of fluency could potentially reflect the first step that feature-involving representational systems take on the path to developing mechanisms for producing propositional representations, transitioning ‘Judge!’ into ultimately ‘I judge that I will walk’ (rather than take the bus or wait it out).

If cognitive affordances index the content of propositional representations, the presence of such an index ought to motivate the possibility that feature-involving systems have a role to play in the origin story of propositional systems. Their indexical relation to content had to become established somehow, and ‘by chance’ is not only unsatisfying, but it could also be incorrect. One plausible proposal for explaining how affordances came to index the content of propositions appeals to how this link became established through the role that feature-involving representational systems play in producing propositional representational systems, a role realized during an individual’s development and enculturation if the social-scaffolding view is on the right track. Of course, to make the case for this proposal, one will need to account for this role and explain its significance in making propositional thought possible. So, let us sketch how this might work according to the interpretation of Proust’s theory of metacognition above.

Proust can be read as proposing that the combination of feature-placing and feature-based representational systems play a crucial role in paving the way for capacities that possess the property of objectivity, distinctive of human cognition. The appeal to a cognitive system with both a feature-placing representational system, which maps functional relationships between bodily acts and features of the distal environment, and a feature-based representational system, which maps functional relationships between cognitive acts and features produced by the cognitive system itself, both of which are systems capable of producing conscious representations at the animal level of description in the form of affordances, could potentially provide a framework for accounting for the origin of the distinction in thought between the subject herself and an objective world. Such a
distinction in thought would then serve as the foundation for the self-world distinction crucial to distinctly-human cognition (see McDowell 1996), i.e. capacities for distinguishing explicitly in thought oneself from the world (and vice versa), and thus the property of objectivity discussed above (see Section 2).

The featuring-placing system, by mapping features in the distal environment, establishes thoughts about the world, while the feature-based system, by mapping features produced by the cognitive system itself, establishes thoughts about the subject herself. Of course, as affordances, these thoughts have impoverished content compared to thoughts constituted by conceptual content. But like propositional thoughts, affordance-based thoughts aim at truth by way of aiming at fluency, and so plausibly reflect thoughts structured by a rudimentary form of objectivity. As such, it is the role that feature-involving systems play in realizing the property of objectivity that explains their role in paving the way for propositional representational systems.

If this is correct, it explains how the content of an affordance (e.g. ‘Judge!’) has come to index the content of a proposition (e.g. ‘I shall walk’): feature-involving systems, by establishing a subject-object distinction in thought, play a crucial role in structuring thought to be about an ‘I’ that engages ‘the world’, thus ensuring the content of an affordance, which calls for action, will represent (in the de re mode) what are often (if not always) central components to any afforded action, the self and the world and any concepts that serve to describe them. Thus together, both feature-placing and feature-based representational systems plausibly comprise essential cognitive components to propositional systems and thus capacities for metarepresentation.

Though several questions remain unresolved (e.g. how this manages to explain the origin of de dicto reference; see Introduction), the plausibility that the origin of the subject-world distinction is rooted in metacognition ought to motivate the idea that metacognition, in line with the role that Proust theorizes it to play, is something over and above, mere ‘run-of-the-mill’ first-order cognition (i.e. plausibly providing a solution to the problem of triviality), describing mechanisms and capacities that might reflect crucial precursors to capacities for metarepresentational metacognition (i.e. possibly making inroads into solving the problem of the origin of self-knowledge). That being said, let us now turn to criticisms of her view, criticisms that were already introduced (see Introduction), and will be discussed in more detail later on (see Chapters 3 and 5).

7.2 Criticisms
Though motivation has been provided above for endorsing Proust’s theory of metacognition as possibly solving the problem of triviality, as well as making inroads into solving the problem of the origin of self-knowledge, there are nonetheless several crucial aspects that remain either unresolved or unclear. In what follows, criticisms of her view will be discussed, and the chapters that feature in-depth treatments of these criticisms will be pointed to.

Consider first whether Proust’s theory, as construed above, can account for differences in metacognitive performance observed across the animal kingdom. This is the problem of metacognitive biodiversity already introduced (see Introduction). Recall that this problem refers to the challenge of accounting for the diverse results from studies in comparative psychology based on the uncertainty monitoring paradigm with various species of animal.

Given the interpretation above, the answer offered by Proust’s theory would likely point to feature-based representational systems that produce cognitive affordances. Here a proponent of her theory could argue that the difference in metacognitive performance among animals lies in whether animals possess the feature-based representational system, such that those that pass the uncertainty monitoring tests do possess this system, while those that fail do not. But there is an issue with this solution. Since feature-based representational systems are distinguished from feature-placing representational systems (the latter of which every animal possesses) by the possession of performance standards for cognitive operations and the capacity to produce cognitive affordances and noetic feelings, which, in turn, explain the animal’s behavior to conform to these standards, failure to pass these tests means either one of three possibilities, and it is unclear which possibility is the correct one.

First, it could mean that these animals do not possess cognitive systems for maintaining performance standards of cognitive operations. Second, it could suggest instead that these animals do not possess cognitive systems capable of producing cognitive affordances. Finally, it could be taken to suggest that these animals do not possess cognitive systems that produces noetic feelings. But from Proust’s theory, it is unclear which is the correct solution to the problem of metacognitive biodiversity. Since to deny the first would be to deny a central means of ensuring animals survive—

15 After all, memorial recall is an example of such a cognitive operation, and denying such a basic capacity from the cognitive repertoire of so many non-human animals (those whose performance in the uncertainty monitoring paradigm is not on par with that of adult humans) would incur the conceptual debt— that no one will wish to pay— of having to explain the impressive mnemonic performance of such animals (e.g. dogs, rats, mice, pigeons) in the absence of any cognitive mechanism for profiling the reliability of operations for mnemonic recall.
one will want to first consider the absence of either cognitive affordances or noetic feelings (or both) for explaining an animal’s failure in uncertainty monitoring tests.

That said, Proust’s theory does not clearly distinguish between cognitive affordances and noetic feelings, and so it is unclear which one of them is absent in the cognitive systems of those animals that fail to pass uncertainty monitoring tasks. Of course, one could point to the recruitment of conscious and/or unconscious bodily arousal to account for noetic feelings as distinct from cognitive affordances. But it is unclear how bodily arousal, if unconscious, would produce distinct behavioral responses, such as those observed in animals utilizing the opt-out mechanism. So, this route would likely wish to pivot to a distinction between unconscious and conscious bodily arousal, such that animals who pass the uncertainty monitoring tasks undergo conscious noetic feelings, which motivate and facilitate their metacognitive performance, while those who fail do not undergo conscious noetic feelings. This route is the one pursued by the present work and shall be revisited later on (see Chapters 2, 4, and 6).

Concerning the route that seeks to account for the problem of metacognitive biodiversity by denying that cognitive affordances are produced by the animal’s cognitive system, this response could collapse into a highly untenable one, so a proponent of this view will need to exercise a due amount of caution. This is because, due to how basic and essential some cognitive operations are for most animals, like operations underlying memorial recall, the animal’s cognitive system is likely to be already in possession of feature-based representational systems for producing cognitive affordances. In light of this, the likely recourse would be to appeal to a distinction between unconscious and conscious cognitive affordances, such that only those animals that exhibit metacognitive performance on par with humans have cognitive systems that produce cognitive affordances in a conscious manner (e.g. through broadcasting them in the global workspace). This route represents an alternative to the one proposed by the present work and will be discussed later on (see Chapters 5 and 6).

In subsequent chapters, I shall argue that an appeal to conscious bodily arousal in the form of conscious noetic feelings of confidence and uncertainty (and the cognitive affordances these feelings present) is the solution to the problem of triviality, as well as its related problems, like that of metacognitive biodiversity (see Chapter 4 and 5). This approach is similar in all but one respect to the alternative approach taken by Meyniel and colleagues (2015a, 2015b). Though their view provides a framework for construing processing fluency signals (and the cognitive affordances these signals would offer) as conscious, it differs from the present account in that it does not appeal to bodily arousal to account for how these signals become conscious (see Chapter 5).
Notice that either approach, i.e. the one taken by the present work or the one taken by Meyniel and colleagues, is quite distinct from the one taken by Proust. While not denying that affordances and noetic feelings can become conscious (see Section 3), her account does not appeal to either conscious cognitive affordances or conscious noetic feelings to shore up metacognition against criticism. Rather, as I interpret her theory above, her approach is to appeal to the role that mechanisms and capacities for metacognition play in providing a plausible origin for a subject-world distinction in thought, and thus a property distinctive of human cognition, specifically the property of objectivity (see Sections 2 and 7.1).

Moving onto other criticisms of her theory, the problem of metacognitive biodiversity gives way to another problem, namely the problem of honeybee metacognition (see Introduction). As already introduced above, Perry and Barron (2013) deployed the uncertainty monitoring paradigm to determine whether honeybees would opt out of hard to discriminate tasks, and the results were quite striking: honeybees performed nearly as well as humans in these tests, suggesting the cognitive systems of honeybees possess mechanisms for metacognition. Clearly, this result has the potential to pose a problem to claims that metacognition ought to be construed as a crucial precursor to capacities for self-knowledge (see Introduction for this notion of cruciality). The possible presence of metacognition in honeybees suggests metacognition is ubiquitous in the animal kingdom, which, if true, would rob metacognition of its privileged status as a crucial ingredient in self-knowledge. As one might readily suspect from the solution to the problem of metacognitive biodiversity, the solution to the problem of honeybee metacognition will likewise involve appealing to consciousness, in particular conscious noetic feelings in the present case and conscious cerebrally-generated feedback signals in Meyniel and colleagues’ case (see Chapter 4 and 5, respectively).

Thirdly, it is unclear how Proust’s theory of metacognition describes mechanisms of cognitive control that are any different from mechanisms that explain basic feats of cognition. Despite the appeal to metacognitive governance, feature-placing and feature-based representational systems, it is nonetheless unclear how the mechanisms for metacognition enable a cognitive flexibility any greater than the cognitive flexibility offered by those mechanisms that explain such basic cognitive feats as adaptive behavior. This is the problem of triviality introduced above (see Introduction), and it will be discussed in detail later on, specifically how Proust’s model for metacognition shares a strong family resemblance to models of reinforcement learning (see Chapter 3).

In a similar vein, Proust’s cybernetic theory of the animal mind shares a family resemblance with theories that emerge from the Bayesian Brain Hypothesis (Knill and Pouget 2004), in particular the theory of predicting processing (Hohwy 2013; Clark 2016), which all describe the
brain as a predictive engine, approximating Bayesian inference in all things cognitive, including perception, action, reasoning, decision-making, and emotional cognition (for more on this, see Chapter 5). As such, this reflects another manifestation of the problem of triviality introduced above (see Introduction), such that metacognition might be rendered as little more than ‘run-of-the-mill’ cognition. So, questions remain about how metacognition enables sophisticated cognitive feats related to reflective and/or executive processes and about what exactly makes metacognition significant so as to be singled out as metacognition.

Given what was discussed above, one might pivot to the feature-based representational system once again, and so argue that cognitive systems that produce cognitive affordances in the form of conscious noetic feelings provide the required tools for establishing the mechanisms of metacognition as sufficiently distinct and sophisticated so as not to be reduced to mechanisms of reinforcement learning or the day-to-day business of cognition. Indeed, this is the route pursued by the present work for solving the problems facing metacognition (see Chapter 2, 4, and 6).

Fourthly, as it concerns the ultimate problem on the origin of self-knowledge, it is unclear how Proust’s theory of metacognition warrants construing it as a crucial precursor to capacities for metarepresentation. Recall from above (see Introduction) that the species of self-knowledge that serves as the present work’s ultimate explanandum is a highly sophisticated form, one that is dependent on the de dicto mode of reference, in which the self is represented as the self. Even though Proust’s theory has been interpreted above as providing a plausible origin for the self-world distinction in thought, it is nonetheless unclear how metacognition, so construed, can explain the origin of representations of the self as the self.

The present augmentation to metacognition, embodied metacognition, attempts to account for de dicto reference by appealing to embodied expressions of confidence and uncertainty that represent (in the de re mode) neurocomputational changes describing features of cognitive performance. It will be argued that these embodied episodes plausibly lay the foundation for both intraindividual as well as interindividual regulatory practices of an epistemic kind, which, if the social-scaffolding view is correct, could plausibly comprise the foundation for a rudimentary form of mindreading and thus a basis for the representation of the self as a self (see Chapters 4 and 6).

Finally, another issue with Proust’s theory, and one which will confront us in the next chapter (see Chapter 2), is that despite its thorough accounting above, it is nonetheless unclear how noetic feelings are actually produced. Current research lacks a comprehensive theory and several conundrums emerge whenever the mechanism of their production is considered. Are noetic feelings produced solely by the brain or is their production dependent upon the extracerebral body
in some crucial way? If their production does turn out to be intertwined with the body, how ought we to account for this relationship? Puzzling are also questions about how noetic feelings are perceived or sensed. What exactly do we feel when we feel noetic feelings? Are we somehow feeling neurocomputational activity or are we sensing bodily activity that has been caused by and/or represents neurocomputational activity (in the \textit{de re} mode of reference)?

Though questions of this sort will be addressed in the next chapter (see \textit{Chapter 2}), other, equally pressing questions await us once inroads have been made into answering the first set of questions. In particular, questions arise about whether noetic feelings are embodied, and, if so, whether their status as embodied plays a crucial role in explaining how noetic feelings fulfil their role in metacognition processing. What is more, questions arise about whether their role as embodied plays a role in providing the required tools for bridging the phylogenetic origin story to the ontogenetic origin story about self-knowledge. In what follows, I answer both of these questions in the affirmative and, in so doing, make the case for embodied metacognition (see \textit{Chapters 4} and \textit{6}, respectively).

\section{Conclusion}

Consider once again the illustration of George, the spider monkey, featured in the \textit{Introduction}, in which noetic feelings of metacognition are contrasted with Reward Prediction Error Signals (RPEs), crucial components in mechanisms for reinforcement learning. Given what was discussed above, Proust will likely want to point to metacognitive governance and the two representational systems of featuring-placing and feature-based systems to shore up her account against strategies that will seek to reduce her theory to one merely about RPEs. In particular, she will want to pivot to how these mechanisms explain the origin of a sensitivity to epistemic norms and how engaging in such behaviors could plausibly give rise to a self-world distinction in thought, thus serving as precursors to capacities for self-knowledge.

But it is unclear how this account can properly address the problem of triviality, specifically made clear by comparing her theory of metacognition to theories of reinforcement learning (see \textit{Chapter 3}), as well as comparing it to the combination of the Bayesian Brain Hypothesis and the theory of predictive processing (see \textit{Chapter 5}). All the while, if the problem of triviality cannot be solved, then there is little hope that it can offer a solution to the ultimate problem on the origin
of self-knowledge (for the present account of how to make inroads into solving this problem by appealing to augmented version of Proust’s theory of metacognition, see Chapter 6)

That said, Proust’s theory of metacognition provides several key resources that will comprise the foundation of the present work’s approach to shoring it up against criticism. Proust teaches us that a proper defense of metacognition should make clear how the underlying metacognitive mechanisms enable enhanced levels of cognitive flexibility. This cognitive flexibility needs to be spelled out in in terms of the wholesale integration of multiple informational sources for the selection of policies regarding cognitive action, which together enable behaviors to conform to norms of an epistemic kind. These metacognitive mechanisms should also play a key role in explaining how subpersonal evaluations and predictions become salient to the subject by way of mental representations at the animal-level of description, potentially as feeling-based affordances.

Instead of pivoting to the role of metacognitive governance in explaining the emergence of epistemic normativity, the argumentative thrust of the present work derives from a construal of noetic feelings as conscious bodily expressions of emotion that enable and facilitate executive control over cognition and behavior. Effectively, this is the proposed solution offered by the present work to the problem of triviality facing metacognition (see Chapter 4).

With respect to the ultimate problem about the origin of self-knowledge, the present work argues that strongly embodied noetic feelings are also the right way to go. This is because of how noetic feeling serve to index proto self-beliefs (see Chapter 4), as well as plausibly enable robust intraindividual and interindividual regulatory practices for coping with uncertainty, which allows bodily expressions of noetic feelings to transform into social-cultural commodities to be detected, monitored, and regulated by other individuals, thus linking the phylogenetic origin story to the ontogenetic origin story about self-knowledge (see Chapter 6).

Effectively, the case for shoring up metacognition is made by arguing throughout the present work that there are reasons to believe that at least some noetic feelings enable and facilitate executive control over cognition and behavior, making it possible for subjects to learn about certainty/uncertainty through emotional experiences characteristic of self-confidence and self-doubt, and thus develop practices around coping with uncertainty, as well as communicating uncertainty within their social-cultural environments. In other words, it will be argued that metacognitive mechanisms of an embodied kind provide us with the crucial tools for building the bridge between the two origin stories and thus for making inroads into solving the problem of the origin of self-knowledge for creatures like us (see Chapter 6).
Chapter 2

Are Noetic Feelings Embodied? The Case for Embodied Metamnemonic Cognition

Abstract

One routinely undergoes a noetic feeling (‘metacognitive feeling’, ‘epistemic feeling’), the so-called ‘feeling of knowing’, whenever trying to recall a person’s name. One feels the name is known despite being unable to recall it. Other similar experiences also fall under this category, e.g. the tip-of-the-tongue experience, the feeling of confidence. Distinguishing noetic feelings from other kinds of feeling is how noetic feelings are crucially related to the facts we know, so much so that the activation of the system of memory theorized to deal in the storage and retrieval of facts, semantic memory, can easily result in the production of a noetic feeling – a regularity that memory research often exploits. And yet, little is known or agreed upon regarding this mechanism. Is it solely brain-based or does it depend upon the extracerebral body in its production of feelings? To arrive at an answer to this question, various studies in metamemory research will be analyzed to determine what ought to be made of the mechanism responsible for noetic feelings. I argue that recent evidence suggests that it relies upon extracerebral processes, in particular cardiovascular processes, the result being support for an embodied view of metamnemonic cognition.

1. Introduction

Imagine having the task of memorizing a list of names for subsequent recollection and being presented with one of these names later on and having to judge whether it was on the previously studied list. One of two mental episodes will typically occur. You might remember the previous experience of memorizing the name or you might look at the name and be struck with a feeling that the name is familiar or known. This is an illustration of the well-corroborated aspect of memory called the ‘remember-know phenomenon’ introduced by Tulving (1985) in support of the distinction between two systems of memory, episodic memory and semantic memory (the former
and the latter cases, respectively). In contemporary memory research, skepticism surrounding the systems theory of memory has emerged more recently (see e.g. De Brigard 2014; Michaelian 2016; Hutto & Peeters 2018; Andonovski 2021). But for the purposes of our discussion, semantic memory (also called ‘declarative memory’) need only serve as a shorthand for those mnemonic mechanisms involved in the recollection of facts, the activation of which is commonly associated with certain noetic feelings, specifically the feeling of knowing or feeling of familiarity (for present purposes, these feelings can be considered equivalent).

Since Tulving (1972) introduced this distinction in his seminal work, episodic and semantic memory have been widely accepted as two distinct and primary forms of memory, so much so that it has been called the ‘standard taxonomy of memory’ (Michaelian & Sutton 2017). Essentially, episodic memory refers to the system that stores and retrieves first-hand experiences, while semantic memory refers to the system that deals in facts about the world. In the philosophy of mind and cognition, much has been said about episodic memory, particularly how best to conceive of it as distinct from semantic memory, by appealing to e.g. the phenomenological property of pastness (Martin 2001). Here, questions about the role of the extracerebral body in episodic memory are generally met with a consensus. This is because the general assumption is that episodic memory is intrinsically tied to emotional experience which, in turn, depends upon the extracerebral body. For example, echoing Kant’s famous declaration, Sousa (2017) writes about episodic memory stating, "Memory, we might say, would be empty without emotion, and emotion blind without memory" (163).

Regarding semantic memory, however, no such consensus exists, and it remains unclear what should be made of the brain-body relationship, if there even be one (going forward, the terms ‘body’ and ‘bodily’ will be used to refer to the extracerebral body). Prominent models of semantic memory do not appeal to feelings, emotions, affective episodes, or bodily processes, often only positing associative semantic networks comprised of various nodes and relationships between nodes in order to account for the mechanisms of semantic memory (e.g. Jacoby et al., 1989; Reder et al., 2000). Meanwhile, Dokic (2014) proposes that the key difference between episodic and semantic memory is not that the former involves affective episodes while the later does not; rather, he argues convincingly, that both involve different kinds of affective episodes, each with its own proprietary kind of noetic feeling.

In what follows, the noetic feelings proprietary to semantic memory will be discussed (except when mentioned explicitly, by ‘noetic feelings’ I mean only those noetic feelings associated
with semantic memory\(^{16}\) and reasons for and against construing the mechanism that produces noetic feelings as embodied will be determined by analyzing the results of seminal metamemory research, as well as the results of recent studies in affective psychology that reveal an intricate relationship between metamemory and interoceptive sensitivity, that is, the capacity to sense and interpret bodily afferents. Essentially, the claim is that any explanation of noetic feelings that leaves out the extracerebral body, particularly the cardiovascular system, foregoes an exhaustive explanation of how noetic feelings fulfill their role in the overarching cognitive economy.

In line with much of the literature on metacognition (e.g. Metcalfe 2000; Koriat 2000; Beran et al. 2012), I will assume for this discussion that noetic feelings are metacognitive, in that their role in the cognitive economy is to guide the unfolding thought process (see Proust 2013). That said, there is an unresolved controversy about whether noetic feelings are properly metacognitive (see e.g. Arango-Muñoz 2018; Carruthers 2016; Greely 2021). This controversy concerns whether noetic feelings are second order, that is, whether they represent internal cognitive states (and so characterize cognitive states about other cognitive states: meta-cognitive) or external features of the world (and so are first order and, thus construed, not meta-cognitive). As such, this debate is somewhat orthogonal to what is claimed here. Even if noetic feelings characterize first-order states, they may nonetheless serve as crucial enablers and facilitators of the unfolding thought process. It is easy to confuse this problem of the representational structure of mental states characterized by noetic feelings with the problem of triviality, but as I shall show later on (in this chapter and Chapter 3), these issues are distinct.

The aim of the present work is to provide strong support for embodied metamnemonic cognition (i.e. to the degree that semantic metamemory depends upon noetic feelings), akin to proposals in the wider family of embodied cognition (e.g. Merleau-Ponty, 1945; Hurley, 1998; Wilson, 2002; Gallagher, 2006; Clark, 2007; Shapiro, 2014). So let me briefly say why I think we should care whether metamnemonic cognition is embodied.

As I understand it for the purpose of this proposal, embodiment is about an extracerebral body being responsible for handling some of the problem-solving work that would otherwise be performed by a central and cerebral process. Essentially, the claim is that the extracerebral body ought to be construed as part of the cognitive process, serving as output, but also responsible for

\(^{16}\) With respect to questions about whether there are embodied noetic feelings associated with episodic memory and whether these feelings are distinct from those associated with semantic memory and, if so, how best to account for these differences, I should say that I will not be addressing any of these question in the present work and any discussion of this will only appear at the conclusion to the present chapter where it is described as a possible avenue for future research. The reason for this is simple: the proposal that episodic memory is associated with embodied noetic feelings strikes me as obviously correct, while the proposal that semantic memory is associated embodied noetic feelings strikes me as far from obvious and so vastly more interesting if true.
Chapter 2: Are Noetic Feelings Embodied?

generating input: “[these are] cases where we confront a recognizably cognitive process, running in some agent, that creates outputs that, recycled as inputs, drive the cognitive process along” (Clark 2007: 185). We ought therefore to care about claims of embodiment, regarding metamnemonic cognition or otherwise, because, if sound, these claims contribute to solving a task already cared about, namely, that of providing an exhaustive explanation of the power of cognition to solve problems.

Before continuing, allow me to define some key terms that will be with us throughout our discussion. As just introduced, a cognitive process is embodied iff an extracerebral process is responsible for doing some of the problem-solving work that would otherwise be performed by a central process. A simple example of this is the role of bodily gestures in the unfolding of thought (Clark, 2008; McNeill 2005): the extracerebral body serves as input to the cognitive process, enabling and facilitating the unfolding of thought and generating components that are cognitive in their own right (see Shapiro 2019). As we shall see, noetic feelings fulfill an analogous role.

From this definition of embodiment, weaker or stronger forms of embodiment can emerge, in which the cognitive process is more or less dependent on the body. As it concerns the proposal that metamemory is embodied, we shall see that different accounts of embodiment are at play in the literature. What I label ‘the weak view’ construes the role of the body as making metamnemonic feedback signals conscious, while ‘the strong view’ sees the body as playing a more significant role in this process, not only making feedback signals conscious, but also making specific metamnemonic information conscious. To be clear, by ‘conscious’, I mean accessible for report (see Introduction).

Thus, the difference between weak and strong embodiment is described below as a matter of the intricateness of the implied brain-body relationship. Specifically, the weak view holds that the brain-body relationship is coarse-grained, such that the body conveys little more than the presence of the metamnemonic signal. Meanwhile, the strong view maintains that the brain-body relationship in metamemory is fine-grained, such that the bodily components not only make metamnemonic signals salient, but also convey metamnemonic information in their own right.

This fine-grained brain-body relationship in metamemory will be argued for by appealing to recent empirical evidence from interoception-based studies on metamemory and noetic feelings. Specifically, these studies provide evidence of a common mechanism between metamemory and interoception (i.e. the capacity to sense and interpret bodily signals across conscious and unconscious levels). Evidence suggests that noetic feelings are causally responsible for making metamnemonic processing fluency signals salient, as well as making features of this fluency salient.
This makes plausible, what I shall call, in the spirit of William James, a ‘Neo-Jamesian theory about noetic feelings’, namely that what you feel, when feeling noetic feelings, are bodily changes (see James 1890; Lange 1887).

I hypothesize that this brain-body relationship obtains due to how changes in metamnemonic processing fluency are indexed to bodily changes. This notion of indexing ought to be read as representing in the de re mode of reference (Quine, 1956; Strawson, 1959), in which strong informational correlations exist between what is doing the representing and what is being represented (see Recanati, 2012). This mode ought to be therefore distinguished from a far more sophisticated mode of reference, de dicto, in which something is represented as what it is (see Keshet & Schwarz 2019). The proposal is thus that certain bodily changes ought to be construed as indexing (de re representing) certain metamnemonic changes.

Before continuing, let’s introduce the map of the contested territory. To the west lies a family of conservative views that can be named the ‘traditional view’ (Section 2). There are two in particular that are prominent in the literature: the ‘direct view’ (Section 2.1) and the ‘indirect view’ (Section 2.2). Both hold that noetic feelings are produced by a mechanism that is solely central and cerebral. Models based on this view will be criticized for how they imply an empirical claim that I argue throughout the latter half of this paper is false.

Meanwhile in the east, a family of liberal upstarts has emerged, emphasizing recent empirical evidence that exposes the need to appeal to extracerebral processes to account for the mechanism that produces noetic feelings. Call this the ‘embodied view’ (Section 3). While investigating the liberal family, two siblings can be discerned: the weak and the strong. After first discussing a condition that would justify claims of strong embodiment (Section 3.1), I discuss first the case for weak embodied cognition (Section 3.2), but I argue in favor of the strong view that maintains a highly intricate brain-body relationship in metamemory (Section 3.3), such that, plausibly, bodily changes serve to index mnemonic changes and, in so doing, convey mnemonic information in their own right (Section 3.4). Thereafter, I conclude with a summary and discuss future research (Section 4).

2. The Problem with Traditional Views of Noetic Feelings

In this section, the problem with traditional views of noetic feelings will be discussed. Essentially, the claim that noetic feelings are non-embodied amounts to making an empirical claim
about their underlying mechanism, and, discussed below (see Section 3), recent evidence suggests this empirical claim is false and so is this claim about their status as non-embodied. First, a direct model of noetic feelings will be introduced and shown to be based on this empirical claim (Section 2.1). Next, this direct model will be provided an additional layer of sophistication, one common in the literature on noetic feelings, in which noetic feelings are mediated by processing fluency, and this more sophisticated, indirect model will then be demonstrated to make the same problematic empirical claim as the direct model (Section 2.2).

2.1 The Traditional Direct Model of Noetic Feelings

As it is construed here, traditional views aim to explain noetic feelings by appealing to semantic networks of various configurations, often in an effort to account for the ‘remember-know phenomenon’ (Tulving, 1985). Mentioned above, this phenomenon is based on an observed discrepancy in mnemonic recall, one which serves to distinguish the semantic from the episodic system of memory. To predict the remember-know phenomenon, Reder et al. (2000) developed a prominent model that distinguishes between two types of mnemonic nodes: a concept node, which holds lexical information (semantic memory), and an event node, which holds information about the encoding event (episodic memory).

As is common in the literature (see Dunlosky & Metcalfe 2008), Reder et al. describe the ‘know’ response as dependent on the subject undergoing noetic feelings and theorize the production of noetic feelings to result from “an elevation in base-level activation” of one or more concept nodes (ibid: 318). Thus, the production of noetic feelings is explained by appealing solely to cerebral mechanisms, such that no explanans appeals to any bodily process. Though this is only one example, more models with a similar explanatory structure, used to elucidate the mechanisms responsible for the production of noetic feelings, will be discussed below.

That said, this claim about the status of the mechanism producing noetic feelings as solely central and cerebral runs counter to contemporary theories in affective psychology about the nature of feelings, and thus this claim is potentially false due to how it implies a claim about the non-embodied status of noetic feelings. Relative consensus surrounds the proposal that feelings are one of the central components to emotion (along with an appraisal, an expression, an autonomic reaction, and an action tendency; see Armony & Vuilleumier 2013). One such increasing influential theory is the Neo-Jamesian interoception-based theory of emotion, which understands feelings as
arising from the internal perception, the *interoception*, of physiological changes occurring within the body (see Critchley & Garfinkel 2017; Barret 2017; Tsakiris & De Preester 2018). Within this framework, feelings are conceptualized as the conscious expression of emotion, so that bodily processes play a crucial role in the cognitive economy, making features of this process salient by producing bodily afferents that are sensed by the subject through interoceptive channels.

Of course, an interoception-based theory of feeling is not the only game in town. Researchers working in metamemory have produced their own theories of feelings. Bower (1981) modifies the influential associative network theory of memory to account for emotions, conceptualizing feelings as memory units that assist in retrieval. Berkowitz (1990) develops the so-called *Cognitive-Neoassociationistic Model* of emotion, in which feelings are elicited by the activation of certain mnemonic nodes. Recently though, theories of feeling in metamemory-based research have struck a different tone. Grandjean et al. (2008) suggest that feelings arise from patterns of neuronal synchronization at different levels of the cognitive hierarchy, but, crucially, their model includes the cardiac rhythm as a peripheral component to emotion. Meanwhile, Thagard and Aubie (2008) argue that feelings are produced by interacting brain areas coordinated in working memory, coordination that occurs through the integration of the perception of bodily states (i.e. via interoception). Thus, as more becomes understood about the brain-body relationship in cognition, models of mechanisms responsible for the production of feelings are changing, from traditional to more embodied accounts, in which the extracerebral body steadily plays a more significant role in the overarching cognitive process.

That said, proponents of the traditional approach might wish to distinguish two distinct species of feeling, noetic feelings and emotional feelings, with the body playing a crucial role only in the production of the emotional sort. In other words, traditionalists might wish to advocate for ‘anti-Jamesian’ species of feeling, a solely cerebral and non-embodied noetic feeling. But such a claim stands in need of empirical support, and the emerging evidence discussed below suggests this position ought to be regarded with skepticism (see Section 3). But before this evidence can be discussed, we ought to examine one key strategy that a proponent of the traditional view might pursue to bolster the case for an anti-Jamesian noetic feeling, one which involves an appeal to metamnemonic processing fluency, a central theme that will be with us throughout our present discussion.
2.1 The Traditional Indirect Model with Processing Fluency

In this section, an approach to defending the traditional model will be discussed, in which the case is made for the plausibility of an anti-Jamesian, solely cerebral, noetic feeling. Though this augmented approach succeeds in accounting for various aspects of noetic feelings, this approach will be interpreted as making the same empirical claim that the unaugmented approach makes above, namely that noetic feeling are non-embodied, and so do not dependent on the extracerebral body for their production, which will be argued below to be false (see Section 3).

Distinct from Reder et al.’s direct model of noetic feelings above, Jacoby et al.’s (1989) widely influential indirect model appeals to processing fluency as an intermediary between noetic feelings and their underlying mechanism. Defined as “the content-independent speed and accuracy of ongoing processing (Topolinski & Strack, 2009a: 1468), so-called ‘metamnemonic processing fluency’ (also called ‘processing fluency’ or just ‘fluency’) is a fundamental component to many prominent models explaining metamemory.

According to this view, fluency plays the role that physiological signals play in Schachter and Singer’s (1962) account of emotional experience: “in the attributional analysis of the feeling of familiarity, the ease with which an idea comes to mind or the relative fluency of accomplishing a task might serve a role similar to physiological arousal in Schachter’s analysis of emotions” (Jacoby et al., p. 395). To better grasp exactly how this appeal has the potential to bolster the traditional view, it will be worth discussing the technical details.

Jacoby and Whitehouse (1989) conducted experiments involving word recognition that consisted of two phases and two conditions, experiments which echo Tulving’s original experiments discussed above. In the first phase, subjects were presented with a list of words to be studied. In the second phase, subjects were presented with a single test-word and had to judge whether it had been on the list. Crucially, sometimes the presentation of a test-word was preceded by the presentation of a context-word, which was either the same as the test-word, different from it, or consisted of a string of nonsense characters (such as ‘&&&&&’). Each experiment involved two conditions: the unaware condition, in which the context-word was visually masked, and the aware condition, in which the duration of the presentation of the context-word was increased and subjects were told about it in advance.

Jacoby and Whitehouse found surprising results. If a masked context-word was the same as the test-word in the unaware condition, subjects were more likely to erroneously report the test-word as previously studied, though this error was not made if subjects were in the aware condition.
Jacoby and Whitehouse stress that these experiments ought not to be construed as demonstrating mere priming effects. Priming is the well-known phenomenon in which the presentation of one stimulus, below the threshold of conscious awareness, enhances the processing of another (Bargh & Chartrand 2014). This is because, highly relevant for our discussion, rather than e.g. demonstrating the increased speed by which subjects recognize primed words, these experiments asked subjects to judge whether a test-word was old or new, which Jacoby and Whitehouse interpret as involving a feeling of familiarity, which itself “rests on an attribution or inference about the source of effects on processing” (ibid: 127).

Jacoby and Whitehouse account for the production of noetic feelings by appealing to specific activation patterns in neural populations (i.e. those occurring in, between, or across concept nodes) theorized to produce metamnemonic fluency signals, which subjects infer, with the aid of a cognitive appraisal, as the familiarity of the stimulus. Thus, noetic feelings are conceptualized as appraised (or evaluated) metamnemonic processing fluency signals. Critically, Jacoby and Whitehouse theorize that fluency signals are produced solely by patterns of neuronal activation, a process that does not depend upon the body, a view about fluency that is still widely held (for a review of processing fluency, see Alter and Oppenheimer, 2009).

If such an account is accepted as sufficiently explaining the mechanism behind the production of noetic feelings, and noetic feelings continue to be conceptualized as feelings in the spirit of affective psychology, the soundness of this account will depend upon a distinction between cerebrally-produced (noetic) feelings, on the one hand, and extracerebral-produced (non-noetic) feelings, on the other. Thus, implicit in this view is an empirical claim about the mechanism behind noetic feelings – an empirical claim which, evidence suggests, is likely false (see Section 3).

Below I argue the traditional view ought to be augmented to model the whole process, including the extracerebral body and its role in producing noetic feelings, as well as the role these feelings play in enabling and facilitating the thought process. In this regard, two avenues for augmentation will be discussed. First, one might provide a convincing case that processing fluency impinges upon a subject’s core affect state, and thus appeal to two central components, one cerebral and one extracerebral, which, when taken together, explain how metamnemonic processing fluency is made a salient to the subject. This is the weak embodied view and it will be discussed first (Section 3.2). Otherwise, one could argue for a stronger brain-body relationship in metamemory by appealing to how bodily afferents not only make metamnemonic processing fluency salient, but how they also make information about this fluency salient (Section 3.3). This is the strong embodied view, and it will be discussed and argued for last (Section 3.4).
3. The Embodied Metamnemonic Cognition Approach

In this section, the embodied view present in the literature on metamemory will be discussed. As I interpret it, proponents of this view, whether weak or strong, maintain a Neo-Jamesian theory about noetic feelings, maintaining that noetic feelings ultimately result from the interoception of bodily afferents. Furthermore, this theory implies an embodied view that extracerebral processes play an essential role in explaining how noetic feelings come to operate in the cognitive economy, serving as both output and input to the thought process. After introducing a condition for strong embodiment (Section 3.1), the weak view is discussed (Section 3.2) prior to the strong view (Section 3.3) and finally I shall argue in favor of the strong view (Section 3.4).

3.1 Condition for Strong Embodiment

Before the findings from emerging research on the relationship between interoception and metamemory can be discussed, it must be clarified what it would mean for a metamnemonic mechanism to be embodied. As already introduced, this involves demonstrating how an extracerebral process is responsible for performing some of the problem-solving work that would otherwise be performed by a central process, thus producing cognitive components in its own right, which serve as both output and input to the overarching process.

To argue for this, I appeal to research in affective psychology, particularly interoception-based research on metamemory. In this respect, it is important to note that the extension of the term ‘interoception’ has been modified recently to become more inclusive. Originally, it referred solely to processes mediated by the viscera and efferent pathways (Sherrington 1906; Langley 1921). But now it describes “the process by which the nervous system senses, interprets, and integrates signals originating from within the body, providing a moment-by-moment mapping of the body’s internal landscape across conscious and unconscious levels” (Khalsa et al. 2018: 501).

Thus, the processing of interoceptive information might potentially supply the basis for any cognitive process that ought to be construed as an embodied one. But one caveat is in order. By definition, interoception is a cerebral process, occurring within the central nervous system (particularly, in the insula cortex), and so appealing to this might sound antithetical to a case for embodied cognition. To avoid this pitfall, bodily afferents will need to be distinguished from the
cerebral processing of those afferents, and these afferents will need to be demonstrated as playing a crucial role in the target cognitive process.

As already mentioned, it is assumed, in line with much of the metacognitive literature, that noetic feelings guide the unfolding thought process. This means, the goal is to show how bodily afferents are responsible for noetic feelings realizing this cognitive role. To argue for this, I interpret empirical evidence that focuses on the role of interoception in metamnemonic cognition, evidence that demonstrates how the production of noetic feelings depends upon an intricate interface between the brain and the body, a common mechanism between interoception and metamemory, in which, plausibly, degrees of metamnemonic processing fluency are indexed to bodily changes and made salient through the interoception of bodily afferents.

To be clear, several views below are nonetheless embodied views. This is because, as I interpret them, these views imply the extracerebral body is an essential part of the problem-solving process due to how changes in metamnemonic processing fluency cause changes in physiological arousal (output), which, in turn, produces noetic feelings that serve as guides (input) to the subject’s thought process. In other words, in order to explain the role of noetic feelings in the overarching cognitive economy, the weak view holds that bodily processes must be appealed to in order to account for how changes in metamnemonic processing fluency ultimately become salient to the subject and thus consciously accessible.

In order to distinguish between weak and strong views, consider this analogy. The relationship obtaining between metamnemonic processing fluency and physiological arousal can be thought of as a process of translation. In this respect, the language of metamnemonic processing fluency is written in patterns of electrochemical activity that describe communication across neuronal populations as to some degree fluently or not fluently processed. If it is correct to say this process is an embodied one, the statements of the source language of cerebral activity will need to be translated, as it were, into the target language of the extracerebral body, i.e. becoming bodily afferents that serve as guides to the thought process.

Intuitively, this process of translation is adaptive because of how emotion reflects one crucial component to nature’s solution to the problem of action, i.e. the problem of what exactly to do now, at this very moment, given so many possibilities (see Zhu & Thagard, 2002; Railton, 2017). The information carried by metamnemonic processing fluency will thus be all the more effective in guiding the subject’s behavior if it is recast in the language of bodily arousal and valence. In this respect, stronger claims about embodiment are those that demand, or advocate for, more faithful translations from the language of brain to that of the body.
Below I interpret the weak view as a claim about a low-fidelity translation, implying a coarse-grained brain-body relationship obtaining between metamnemonic processing fluency and bodily afferents (Section 3.2). In particular, the weak view appeals to bodily afferents to explain how metamnemonic processing fluency becomes salient to the subject, though it is unclear whether the studies below provide evidence to support this view. Roughly speaking, proponents of the weak view construe the body as a kind of signal booster, amping up processing fluency signals to become accessible for conscious report, but they stop short of appealing to an extracerebral process to explain how degrees of metamnemonic processing fluency become salient to the subject.

Meanwhile, the strong view is a claim about a high-fidelity translation, which implies a fine-grained relationship obtains between the brain and the body in metamnemonic cognition (Section 3.3). Here, degrees of fluency can be made salient to the subject by becoming translated into bodily arousal and valence, which the subject senses through interoceptive channels, due to how metamemory and interoception are theorized to be underlaid by a common mechanism.

I argue in favor of the strong view by discussing how the empirical evidence points in this direction, suggesting that bodily afferents serve to index features of metamnemonic processing fluency (Section 3.4). Of course, if either the weak or the strong view is true, the traditional view, which construes the mechanism responsible for producing noetic feelings as independent of the extracerebral body, is false. So, in building the case for either embodied view, I am also making the case that the traditional view should be regarded with a due amount of skepticism.

3.2 Weak Embodied Metamnemonic Cognition

According to the weak view, noetic feelings are caused by the interoception of bodily afferents. But this view falls short of articulating a fine-grained relationship between the brain and the body in metamemory. As will be discussed, some weak views collapse back into traditional ones, while others simply fail to hold a position that can be characterized as strong. Somewhat surprisingly, though the views below are embodied views of metamemory, the studies upon which those views are based struggle to provide evidence for holding any variant of an embodied view. Such evidence will not be arrived at until the next section when the strong view is discussed (Section 3.3). Below the weakest view is discussed first, followed by a discussion of progressively stronger views.

The first instance of an embodied view comes from a study conducted by Goldinger and Hansen (2005) that found that a subliminal buzzing would enhance feelings of familiarity, showing
that subjects under the buzzing’s influence were more likely to report a stimulus as ‘old’ both correctly and incorrectly (for a discussion of this study and similar studies with respect to the debate about the metacognitive status of noetic feelings, see Arango-Muñoz, 2019). Crucially, the buzz not only increased confidence in false alarms, but, moreover, it reduced confidence in true positives. To make sense of this, Goldinger and Hansen appeal to Whittlesea and Williams’ (2001: 3) discrepancy-attribution hypothesis: “When the quality of processing is perceived as being discrepant from that which could be expected, people engage in an [inferential] process.”

Accordingly, the physiological arousal generated by the subliminal buzzing produces differential effects depending on subjects’ expectation. If the stimulus is new, the heightened sense of familiarity spurred on by the subliminal buzzing is inferred to suggest the stimulus has already been encountered, explaining the increase in false positives. However, if the stimulus is old and the feeling of familiarity is extrinsically heightened, it becomes inferred as artificial, raising the specter of doubt in the mind of the subject about whether it is indeed familiar, thus explaining the reduction in confidence of true positives. Goldinger and Hansen believe this suggests subjects can, in the best-case scenario, rely on a signal distinct from physiological arousal to guide their thinking process: a metamnemonic processing fluency signal issued by the process of encoding and retrieval that reflects “internal processing changes” (ibid: 526)

Thus, according to Goldinger and Hansen, physiological arousal is one of the various cues subjects utilize while appraising stimuli as familiar, so long as the metamnemonic processing fluency signal alone is insufficient for this appraisal. Though bodily afferents (in the form of proprioceptive vibrations) are demonstrated to exercise a causal influence on the production of noetic feelings, what Goldinger and Hansen construe as “gut feelings” are highly susceptible and at times even unnecessary for the overarching cognitive process. This view is the weakest variant of the embodied views, approaching even traditional views, since Goldinger and Hansen do not envision the bodily afferents as playing any integral role in the thinking process except when the metamnemonic signal fails to fulfill its role in the process.

A similar study conducted by Allen et al. (2016) measured the influence of unexpected arousal on confidence reports. They did this by presenting masked disgust-cues in advance of a visual discrimination task of variable sensory precision: stimuli were either precise (and so easy to discriminate) or blurry (and so hard to discriminate). Similar to the study above, the subliminal disgust cues decreased confidence for visually precise trials but increased confidence in noisy ones.

Allen and colleagues explain this inverted relationship differently. Appealing to the Bayesian Brain Hypothesis (Knill and Puget 2004) and the predictive processing framework (Hohwy 2013;
Clark, 2016), Allen et al. interpret the unexpected arousal as counteracting the influence of cognitive processing biases (e.g. biases for perceptual motion, such as how noisy objects appear to move slower; see Vintch & Gardner 2014). Such biases are theorized to be computational priors within the cognitive hierarchy that modulate the influence of sensory information on experience.

If this is correct, Goldinger and Hansen’s study could be reinterpreted to show how subliminal buzzing serves to counteract priors about mnemonic processing: under the influence of the subliminal buzz, higher fluency is experienced as indicating an unfamiliar stimulus, while lower fluency indicates a familiar stimulus, a result which Allen and colleagues see as “motivat[ing] a revised view of metacognition as incorporating [priors] about both physiological states and the precision of actual sensory inputs” (ibid: 7).

In other words, applying Allen at el.’s gloss to this study yields the proposal that before noetic feelings come to guide the thinking process, the cognitive system will first draw upon computational priors concerning not only patterns of mnemonic activity but also patterns of bodily activity. Bodily afferents are thus construed as integral to this cognitive process of determining familiarity, confidence, etc. However, it remains nonetheless unclear whether, in Goldinger and Hansen’s study, the interoception of bodily afferents explains the process by which metamnemonic processing fluency signals become salient to the subject, let alone whether the bodily afferents make information about these signals salient. So even after applying Allen et al.’s embodiment-friendly gloss, Goldinger and Hansen’s study struggles to provide us with evidence for holding even a weak embodied view of metamemory.

Progressing toward stronger views, Morris et al. (2008) measured subjects’ skin conductance responses (SCR) while subjects reported either remembering or knowing words that had been either previously studied or not previously studied (see the remember-know phenomenon above). They found interesting results: “…SCR latencies were significantly longer in response to studied words than in response to non-studied words. In addition, both studied and non-studied words were given higher recognition ratings when they were associated with long latency than when they were associated with short-latency SCRs” (ibid: 1384).

Due to how previous studies have shown that SCR latencies increase during attention-demanding tasks (see Dawson et al., 2000), Morris et al. interpret these results as providing evidence that the feeling of familiarity “stem[s] from autonomic arousal associated with cognitive resource allocation” (ibid, p. 1378). This comprises our first genuine embodied view: essentially, these researchers propose that noetic feelings are products of the autonomic arousal generated whenever the central nervous system requires the allocation of additional resources.
According to Morris and colleagues’ view, bodily afferents become inferred by the subject with the aid of a cognitive appraisal as indicating an epistemic value, e.g. familiarity, knowledge, etc. (ibid, p. 1379). As such, bodily afferents are responsible for making the allocation of additional resources salient to the subject. Thus, their view ought to be categorized as an embodied view, one which could potentially avoid the pitfalls of making a false claim about two distinct kinds of feeling, one solely-cerebral and another extracerebral kind.

Though, Morris et al.’s view is a stronger than that of Goldinger and Hansen, as bodily afferents are construed as making metamnemonic fluency salient, their study stops short of providing evidence of this. This is because of how SCR latencies were only found to correlate significantly with two features. First, the length of latency correlated with the study status of the item: longer for studied and shorter for non-studied items. Second, longer latencies correlated with reports of highly familiar stimuli, while short latencies correlated with reports of less familiar stimuli. Though this suggests an interesting link between bodily afferents and mnemonic processing, it is not evidence for an embodied view of metamemory, since it is nonetheless unclear whether these latencies are sensed by the subject and guide her unfolding thought process, i.e. whether these afferents explain how metamnemonic fluency becomes consciously accessible. In other words, it is unclear whether this is evidence of causation or whether this is merely evidence of correlation. Though Morris and colleagues hold an embodied view, their study struggles to provide evidence for either a coarse-grained or fine-grained relationship obtaining between the brain and the body in metamemory.

Moving on to a stronger view, Köhler’s memory lab investigated the relationship between the feeling of familiarity and cardiovascular afferents in two experiments (Fiacconi et al. 2016; Fiacconi et al. 2017). The first experiment will be discussed here, while the second will be discussed in the next section. In the first experiment, subjects were given a facial recognition task while cardiovascular signals were monitored via electrocardiography. This enabled researchers to synchronize the presentation of memory probes to the two phases of the cardiac cycle: systole, when cardiovascular feedback is the strongest, and diastole, when feedback is the weakest.

Fiacconi et al. (2016) were able to obtain highly interesting results, namely that cardiovascular feedback can exercise an influence on metamnemonic judgments: faces presented during systole were more likely to be judged as old than faces presented during diastole (regardless of whether the faces were targets or lures, or whether those faces were emotionally laden or neutral in countenance). Crucial to our discussion, this correlation was found only to hold for the ‘know’ response, with its associated noetic feelings, suggesting this informs about the mechanism underlying the production of noetic feelings (see ibid: 567).
Though this experiment provides further evidence of an embodied view of metamemory, as cardiovascular feedback could be construed as a somatic marker indicating familiarity (for somatic markers, see Damasio 1996), these results, for the same reasons as above, fail to provide sufficient evidence for holding either a weak or strong view. Even though the production of cardiovascular feedback is predictive of an increased likelihood that subjects will report a stimulus as familiar, it is nonetheless unclear whether this shows that the cardiovascular signals play any crucial role in explaining the unfolding thought process. This could nonetheless be mere correlation. That said, the results of the second experiment conducted by Köhler’s lab will be interpreted alongside these to provide evidence for holding a strong view (see Section 3.3).

Let us turn to one final weak view, a prominent theory that captures the commonalities between the weak views. Topolinski and Strack (2009b) explain the mechanism behind noetic feelings with reference to semantic coherence, a measure of the degree to which disparate words form a semantic whole. For example, taken together ‘salt’, ‘deep’, and ‘foam’ will imply the sea. Their theory shares crucial features with traditional models already discussed: memory is likewise conceived of as “a network of nodes connected by links of varying strengths” which “consists of cognitive units (e.g. propositions) encoding various facts” (Anderson 1983: 267). But with respect to the mechanism that produces noetic feelings, Topolinski and Strack believe changes in fluency can trigger an affective response which, when taken together, constitutes the process by which noetic feelings are produced.

In this respect, Topolinski and Strack refer to Russell’s (2003) notion of ‘core affect’, suggesting bodily afferents form an essential part of the process through which noetic feelings are produced: “This fluency impinges on the current affective state, which can be understood as “core affect”… [which] is a diffuse and automatic assessment of hedonically important factors such as the physiological milieu, but also reflects all the information processing going on” (Topolinski & Strack 2009a: 1469). Thus, Topolinski and Strack maintain that it is in virtue of changes in physiological arousal that fluency signals become salient to the subject, theorizing that the mechanism underpinning noetic feelings is embodied due to how the body serves as both output and input to the cognitive process: the relevant bodily afferents give fluency signals their affective or experiential character, making these signals salient to the subject in the form of noetic feelings which guide the thinking process.

Though this view is indeed an embodied one, it is unclear whether it makes claims about a coarse-grained or a fine-grained brain-body relationship in metamemory. On the one hand, their view suggests fluency signals can impinge upon the subject’s affective state and cause the production of bodily afferents, which will, in turn, shape the thinking process. But it is unclear
whether bodily afferent satisfy a role over and above that of a signal booster, simply making the fluency signal salient, as though ‘turning up the gain’ on the signal.

In short, a great deal hinges on what exactly it means for the current affective state to “reflect all the information processing going on” (ibidem). This could be read as implying a stronger claim about how certain features of fluency are reflected, or represented, by bodily afferents. If this construal is correct, it would suggest Topolinski and Strack advocate a strong view about the embodiment of metamemory. But, as it stands, the wording is too vague to classify with certainty.

Though several views were discussed that describe the role of bodily afferents in making fluency signals salient to the subject, no empirical evidence of this relationship in metamemory has been provided as of yet. In the next section, not only will evidence be discussed for claiming that bodily afferents cause metamnemonic processing fluency to become salient, but also evidence for claiming that, in so doing, bodily afferents represent changes in metamnemonic processing fluency and so make degrees of fluency salient.

### 3.3 Strong Embodied Metamnemonic Cognition

From our discussion above, a theory about the mechanism responsible for noetic feelings has emerged. The central nervous system, equipped with an associative semantic network, manipulates stored and concurrent information, for the purpose of facilitating some task-specific goal, at times generating an evaluation/prediction of this processing, which, in turn, triggers some degree of physiological arousal that causes this evaluation/prediction to cross the threshold of conscious awareness, at which point it becomes interpreted by the subject with the aid of a cognitive appraisal to imply an epistemic value, e.g. familiarity, knowledge, etc.

A strong embodied account will need to argue that bodily afferents are not only integral to this process (as they are believed to be above), but also that the relationship between the brain and the body is fine-grained, such that bodily afferents not only play a role in making the processing fluency signals salient, but also are responsible for making information about the signals salient. In what follows, recent empirical evidence is interpreted to suggest a fine-grained relationship holds between the brain and the body in metamemory. Ultimately, I shall propose in the next section a hypothesis regarding this mechanism: plausibly, the mechanism ensures cardiovascular afferents represent changes in metamnemonic processing fluency through a quasi-translational process of indexing. As before, I discuss progressively stronger cases.
On the basis of recent neuroimaging studies that provide evidence that both metamemory (e.g. Chua et al. 2006) and interoception (e.g. Critchley et al. 2002) recruit the insula region, Chua and Bliss-Moreau (2016) investigated whether metamemory and interoception are underpinned by a common mechanism. They found “interoceptive accuracy and metamemory accuracy […] were related such that individuals with higher interoceptive accuracy also had better [metamemory accuracy]” (ibid: 155).

Using a combination of Schandry’s (1981) heartbeat perception task and a facial recognition task (Minear & Park, 2004), Chua and Bliss-Moreau compared subjects’ ability to perceive their own heartbeat accurately to their ability to form accurate metamnmonic judgments. After a study phase, subjects paired names with faces and then were asked to rate their confidence in having chosen the correct name. This is widely referred to as a ‘judgement of learning’ (JOL; e.g. Koriat 2000). Effectively, Chua and Bliss-Moreau’s study provides evidence that the better subjects are at accurately detecting their own heartbeat, the more successful will subjects be in making accurate evaluations of their mnemonic performance which suggests “both processes rely on a common mechanism” (ibid: 156).

Details surrounding this study will be discussed below, but before doing so, one crucial component of this study (and studies below) needs to be clarified. Obviously, JOLs are members of a distinct species of mental entity than that of noetic feelings (after all, judgements are not feelings). However, it is safe to interpret, once certain conditions are met, that studies measuring JOLs inform about the mechanism behind the production of noetic feelings. Researchers have argued convincingly that both the casting of JOLs and the production of noetic feelings ultimately rely upon metamnmonic processing fluency signals and the familiarity of context-relevant cues, resulting in the well-corroborated hypothesis, ‘the cue-familiarity heuristic’ (see Metcalfe 1993; Schwartz 1994).

On the strength of this hypothesis then, one can infer the results of any study designed to inform about the mechanism underpinning JOLs as also informing about the mechanism behind noetic feelings, so long as noetic feelings, which motivate and serve as the basis for the JOLs, be distinguished from the JOLs themselves. In our discussion, I focus on the process that motivates the casting of the JOL, which can be safely assumed to involve a noetic feeling.

Returning to Chua and Bliss-Moreau’s theory of noetic feelings, they propose a common mechanism underpinning the capacity to reliably detect bodily afferents, called ‘interoceptive sensitivity’, and the capacity to cast reliable JOLs. If JOLs are made on the basis of noetic feelings, as the cue-familiarity heuristic predicts, it follows that Chua and Bliss-Moreau also hold that the
mechanism enabling the sensing of bodily afferents is either the same mechanism or shares crucial cognitive components with the mechanism that enables the sensing of noetic feelings, so that the mechanism behind noetic feelings ought to be construed as crucially tied to interoception.

It is less clear, however, whether their study provides evidence that subjects are sensing bodily afferents while undergoing noetic feelings. But since performance in the interoceptive domain predicts performance in the metamemory domain, the results from Chua and Bliss-Moreau’s study do begin to make a case for this claim of embodiment. Plausibly, certain subjects perform better than others in tasks of metamemory in virtue of the fact that their increased interoceptive sensitivity provides them with increased sensitivity to noetic feelings, which generally serve as good guides to the unfolding thought process. But without further evidence it is unclear whether this finding suggests causation over mere correlation.

Thus, the crucial question becomes whether such correlations are explained by a common mechanism. If this is correct, their claim of embodiment will be justified, specifically the claim about how processing fluency signals become salient to the subject through the interoception of bodily afferents, suggesting a coarse-grained brain-body relationship in metamemory, possibly even a fine-grained one. In what remains, the case for this common mechanism (and thus embodied metamnemonic cognition as a whole) will be made.

In making the case for embodied metamnemonic cognition, let us consider the second study conducted by Köhler’s lab which measured whether interoceptive sensitivity moderates “the relationship between task-related cardiovascular changes and [feeling of knowing (FOK)] ratings” (Fiacconi et al. 2017: 72). FOK ratings measure the degree to which stimuli are perceived by subjects as familiar (as more or less familiar looking). Similar to casting JOLs, rating a FOK is theorized to occur on the basis of undergoing the noetic feeling (see Koriat and Levy-Sadot 2001).

After measuring interoceptive sensitivity via Schandry’s heartbeat perception task, Fiacconi et al. measured task-related cardiovascular changes while subjects provided forced FOK ratings in facial recognition tests. Fiacconi et al. observed a positive relationship between heart rate acceleration and the reported strength of the FOK, a relationship modulated by interoceptive sensitivity: “Together, these results suggest that relative heartbeat acceleration to old face cues is associated with stronger subjective feelings of knowing for old as compared to new face cues in individuals with high interoceptive sensitivity” (ibid, p. 75). Crucially, when analysis was restricted to know-response, in which subjects report undergoing noetic feelings, “the relationship between cardiovascular responses to the face cues and FOK ratings” was found to be “moderated by participants’ interoceptive sensitivity” (ibidem).
Thus, Fiacconi and colleagues found that for those cases in which subjects could only rely on their noetic feelings, the degree to which subjects could sense bodily afferents determined the degree to which heartrate acceleration and FOK ratings correlated. The better subjects are at sensing bodily afferents, the more likely changes in heartrate acceleration will match the reported strength of the FOK, such that greater/weaker heartrate acceleration means stronger/weaker reported intensity of noetic feelings.

In other words, the reported degree of arousal of the noetic feeling was found to correlate with the degree of heartrate acceleration, a relationship modulated by interoceptive sensitivity, suggesting, “to the extent that participants can ‘tune into’ visceral feedback, autonomic signals do indeed shape FOK judgments” (ibid: 77). Though, they are not explicit about what exactly it means to ‘shape’ both the feeling and the judgement in this manner, some specifics can be inferred by combining the results of both experiments.

The combination of both results (i.e. from Fiacconi et al. (2016) and Fiacconi et al. (2017)) suggests a strong embodied view, in which bodily afferents not only make metamnmonic processing fluency salient, but also make information about this fluency salient, namely degrees of fluency. Thus, evidence for the strong view is provided by combining two findings. On the one hand, the onset of noetic feelings correlates with the production of cardiovascular afferents, while, on the other hand, the reported strength of noetic feelings is a function of both heartrate acceleration and the degree to which subjects can reliably sense the signals that convey this acceleration. Together these results motivate adopting a Neo-Jamesian theory about noetic feelings and a strong embodied view about metamnmonic cognition, both of which will be defended below.\textsuperscript{17}

### 3.4 A Neo-Jamesian Theory of Noetic Feelings

Let us discuss whether the above results provide evidence for a Neo-Jamesian theory about noetic feelings, i.e. that what you sense while undergoing noetic feelings are ultimately changes in the extracerebral body. Though more research of both a conceptual and an empirical nature will need to be done to secure the case for a common mechanism between metamemory and

\textsuperscript{17} A study by Dunn et al. (2010) found a similar moderating effect of interoceptive sensitivity on the relationship between bodily reactions and cognitive-affective processing during tasks of intuitive decision-making, which is often theorized to depend upon noetic feelings (see Dunlosky and Metcalfe 2008).
interoception, the combined results from these studies above provide a strong empirical basis for the case for strongly embodied metamnemonic cognition, and in so doing, also support adopting a Neo-Jamesian theory about noetic feelings. Let’s put the pieces together.

On the basis of Fiacconi et al. (2016), we can infer that a noetic feeling can be produced, even erroneously (i.e. even for lures) if the presentation of a stimulus is synchronized with the systolic phase of the cardiac cycle, in which cardiovascular feedback is strongest. Meanwhile, this relationship is not observed if stimuli are presented during the diastolic phase of the cardiac cycle, in which cardiovascular feedback is weakest.

Of course, Goldinger and Hansen (2005) found that subliminal buzzing could also produce noetic feelings, whose influence on behavior was counteracted when subjects were told about the buzzing, a result which Goldinger and Hansen believe showed how subjects have access to a genuine signal to guide the thinking process. That said, their results do not show that the genuine signal is solely cerebral. It only suggests there is a normative signal, one which subjects generally have access to. Nothing about these results rules out the possibility that the normative signal is an embodied one, produced by a common mechanism between metamemory and interoception. Perhaps, as implied by Fiacconi et al. (2016)’s results, subjects can access a cardiovascular signal that conveys information about metamnemonic processing fluency.

Let’s now return to Fiacconi et al.’s (2017) study. Here, an increase in heartrate acceleration was found to correlate with the onset of noetic feelings, and not with mnemonic recall itself. Moreover, the degree of acceleration was found to correlate with the reported intensity of the feeling, a relationship moderated by interoceptive sensitivity. Though this fact is not mentioned, greater heartrate acceleration increases the probability that cardiovascular afferents will be sensed, even by those subjects who demonstrate poor interoceptive sensitivity (see Garfinkel et al. 2013). Ultimately, the degree to which subjects are sensitive to bodily afferents determines the degree to which heartrate acceleration predicts the reported strength of the noetic feeling: greater/weaker acceleration correlates with greater/weaker reported intensity of the noetic feeling.

This crucial finding points the way beyond mere correlation to the possibility of causation, since it suggests that the better one is at sensing or tuning into cardiovascular afferents, the more intense will be the noetic feelings, a relationship explained by how these afferents could cause these feelings. Thus, these results provide evidence for a Neo-Jamesian theory of noetic feelings, since this study demonstrates that the more sensitive you are to cardiovascular afferents, the more intense will be your noetic feelings in the midst of these afferents being produced.
That said, as previously discussed, Morris et al. (2008) observed similar patterns with SCRs, such that greater/weaker SCRs were found to correlate with greater/weaker intensity of noetic feelings, which was argued above not to evidence claims of embodiment. But the crucial difference in Fiacconi et al.’s studies lies in how cardiovascular afferents are commonly accessible for report: subjects routinely feel their own heartbeats (from the inside), especially if heartrate is accelerating. Meanwhile, the electrodermal activity of SCRs measured by Morris et al., though often a marker of conscious activity, is not itself theorized to be consciously accessible (see Dawson et al. 2000). Since across both studies, Köhler’s lab measured the relationship between, on the one hand, consciously accessible afferents and, on the other hand, the onset and intensity of noetic feelings during the production of those afferent, their results support a Neo-Jamesian theory of noetic feelings.

Consequently, it is plausible that bodily afferents are causally responsible for making mnemonic processing fluency salient, thus supporting for case for weak embodiment, as bodily afferents guide the unfolding thought process, analogous to how bodily gestures can. What is more, these results suggest cardiovascular afferents convey information about the degree to which stimuli are fluently processed, thus lending support to the case for strong embodiment, since subjects “tune into” their bodily afferents to discover the degree to which certain stimuli are fluently processed.

As a result, these studies provide evidence in favor of Chau and Bliss-Moreau’s proposal of a common mechanism shared by interoception and metamemory, realized through the production of embodied noetic feelings that guide the unfolding thought process. Specifically, capacities for metamemory and interoception share a common mechanism due to how the process by which noetic feelings fulfill their role in the cognitive process is enabled and facilitated by the interoception of bodily afferents, possibly those afferents produced by the cardiovascular system.

One last question will need to be addressed before our discussion can conclude. Given what was discussed above about the intricate brain-body relationship in metamemory, a proposal is needed for how bodily afferents are able to convey degrees by which stimuli are fluently processed. For this I submit the following hypothesis. One proposal that strikes me as deeply plausible is that mental states characterized by noetic feelings might represent changes in metamnemonic processing fluency. Specifically, bodily arousal and valence might serve to index crucial processing features: degrees of bodily arousal could index degrees of fluency, while valence could index whether stimuli are fluently or not fluently processed. If this is correct, the function of the common mechanism between metamemory and interoception is not unlike that of an interpreter, translating the language of the brain into that of the body, plausibly offering a high-fidelity rendition for subjects with healthy heart-brain relationships.
Finally, it is clear from our discussion that traditional theories ought to be looked upon with a due amount of skepticism, while embodied theories ought to pave the path forward. This is because of how the traditional theories rely upon an empirical claim that is likely false. Recent empirical evidence suggests noetic feelings are not too terribly distinct from other kinds of feelings, since noetic feelings depend upon extracerebral processes for their production and the interoception of bodily afferents for their realization within the cognitive economy. Though a complete theory of the mechanism underpinning interoception and metamemory currently lies outside our reach, it is clear from the emerging interoception-based research in metamemory that noetic feelings depend upon a combination of both metamnemonic and interoceptive mechanisms, so much so that it suggests adopting a Neo-Jamesian theory about noetic feelings and a strong embodied view of metamnemonic cognition.

4. Conclusion and Further Research

The chief question investigated here was whether there are grounds for claiming metamnemonic cognition is embodied in some crucial way. This question was answered in the affirmative as recent empirical evidence suggests that bodily afferents explain how metamnemonic processing fluency is made salient to the subject, as well as explain how degrees of metamnemonic processing fluency are made salient. In particular, cardiovascular afferents were suggested as playing this role. These results suggest the traditional view should be looked upon with a due amount of skepticism and a strong embodied view that endorses a Neo-Jamesian theory about the nature of noetic feelings should be preferred: when feeling a noetic feeling, you are feeling bodily changes.

One important direction of research opened up by the framework outlined above would seek to determine more details surrounding the common mechanism shared by capacities for interoception and metamemory. This could potentially involve exploring the apparent relationship between noetic feelings, agency, and the cardiovascular system. Consider Thayer and Lane’s prominent neurovisceral integration model (NVM) that conceptualizes changes in heart rate variability (HRV) as “an index of activity in a set of neural structures involved in physiologic, affective, and cognitive regulation” (2009: 86). Also consider Smith et al. (2017) recent update to NVM based on the Bayesian Brain Hypothesis, in which greater HRV reflects executive levels of control. Given what was discussed above about the link between noetic feelings and cardiovascular feedback, might NVM have the tools to describe the common mechanism between interoception and metamemory? Could changes in HRV predict processing fluency effects on behavior, and
might these effects be modulated by interoceptive sensitivity? Addressing these questions will help shed light on the proposed common mechanism between interoception and metamemory (see Chapters 4 and 6).

Second, we might inquire into the embodied nature of other noetic feelings. For example, what are the details surrounding the relationship between interoceptive sensitivity and episodic noetic feelings, e.g. the feeling of mental time travel? Could these noetic feelings have their own unique bodily basis? Meanwhile, our discussion focused on the feeling of knowing/familiarity, but there are more semantic noetic feelings, each with distinct phenomenal character. How might these differences be accounted for? Unfortunately, the semantic variant of noetic feelings are all we have time for in the present dissertation.

Third, there are also several deeply profound questions. Why does the cardiovascular system play a fundamental role in enabling and facilitating advanced cognitive capacities, like capacities for discrimination, identification, and recollection? Is this mechanism online at birth or does it develop throughout childhood? Might there be evidence of a similar brain-body metamnemonic mechanism in other animals? And if so, does this mean that these animals experience noetic feelings? Do noetic feelings, qua feelings, play an important role in the phylogenetic origin of distinctly human forms of self-knowledge (see Chapters 4, 6, and Conclusions)? To address these issues properly, much philosophical and empirical work will need to be done.
Chapter 3
Problems Facing the Defense of Evaluative Metacognition:
Shea’s Metarepresentation Approach and Proust’s Cognitive Role Approach

Abstract

A debate rages on between evaluativists and metarepresentationalists about whether evaluative metacognition (henceforth, ‘metacognition’) is sufficiently robust to qualify as properly metacognitive (e.g. Carruthers 2009a, 2017, 2020; Proust 2009, 2013). Metarepresentationalists charge metacognition as trivial, as arising from cognitively inexpensive mechanisms that render it mere first-order cognition (non-meta) and rob it of its significance in explaining the origin of self-knowledge (i.e. the problem of triviality and the problem of the origin of self-knowledge; see Introduction). In this chapter, two possible approaches to defending metacognition against criticism will be discussed with the goal of diagnosing the reasons why these approaches fail, or would fail, in the eyes of skeptics.

The first approach, dubbed ‘the metarepresentation approach’, amounts to construing noetic feelings as mental representations with a peculiar brand of metarepresentational structure, which Shea (2014) argues reward prediction error signals (RPEs) possess. This approach will be demonstrated to have poor prospects for defending metacognition due to how it would fail to secure a metarepresentational reading for noetic feelings in the crucial mindreading sense (see Introduction). The second approach is the canonical one advocated by Proust (2013), in which the defense of metacognition is made by appealing to the cognitive role that noetic feelings play in enabling and facilitating sophisticated forms of control over cognition and behavior. However, Proust’s double accumulator model for metacognition will be shown to share such a strong family resemblance to the actor-critic model, which describes the role of RPEs in reinforcement learning, that it remains unclear how this approach could properly shore up metacognition against criticism. After a discussion of how best to distinguish noetic feelings from RPEs, it will demonstrated how this discussion points the way toward appealing to noetic feelings as conscious. In the next chapter,
the problem of triviality will be revisited, and there I shall argue it can be solved by augmenting evaluative metacognition to be embodied metacognition (see Chapter 4).

1. Introduction

This chapter has the goal of demonstrating how two approaches, one hypothetical and one canonical, fail to shore up evaluative metacognition (henceforth, ‘metacognition’) against the problem of triviality, and in failing to do so, also fail to provide the necessary tools for building a bridge between the phylogenetic and ontogenetic origin stories about self-knowledge (see Introduction). The first approach is inspired by Shea’s (2014) construal of reward prediction error signals (RPEs) in reinforcement learning as featuring a peculiar brand of metarepresentational structure. Essentially, one might attempt to defend metacognition against criticism by applying this construal to the products of metacognition (noetic feelings) in an effort to argue for their status as metarepresentational in the crucial mindreading sense. The second approach is the canonical one defended by Proust (2013) that involves appealing to the cognitive role that noetic feelings play in enabling and facilitating sophisticated forms of control over cognition and behavior. Both approaches will be demonstrated to encounter issues when attempting to defend metacognition against the problem of triviality.

To foster internal consistency to the present work, I shall also like to bring our discussion in Chapter 2 about metamemory into the broader context of metacognition. To do this, I shall revisit the taxonomy of the traditional and embodied views and demonstrate how the metarepresentation approach can be categorized as a traditional view, while the cognitive role approach can be categorized as a weak embodied view. Meanwhile, I shall defend a strong embodied view in the next chapter (see Chapter 4).

Thus, I shall demonstrate how this taxonomy applies equally well in the literature on metacognition concerning views of mental representations that are argued to be metacognitive (or could theoretically be). But their status as metacognitive is still under debate, so I shall often refer to these representations as ‘candidate signals’. Accordingly, I shall argue in this chapter that the approach taken by either the traditional view or the weak embodied view will only perpetuate the stalemate between proponents and opponents of metacognition (see Introduction), and this taxonomy will be revisited again once an alternative account to the one proposed by the present work is discussed (see Chapters 5 and 6).
That said, the problem with these two approaches discussed in this chapter is not so much that they are traditional or weakly embodied, but rather how they struggle to solve the problem of triviality. In the next chapter, this problem will be solved by appealing to strongly embodied noetic feelings (Chapter 4), while in the chapter thereafter, the problem will be solved by appealing to non-embodied conscious confidence signals (Chapter 5). However, in the final chapter an appeal to embodiment will be argued to better explain recent empirical evidence from interoception-based studies on metacognition discussed above and below (Chapters 2 and 4), as well as provide a more nuanced array of tools for building the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge (Chapter 6). So, an appeal to embodiment will turn out to be advantageous with respect to accounting for this empirical evidence, as well as making inroads into solving the problem of the origin of self-knowledge (but not necessarily advantageous for solving the problem of triviality).

This chapter is divided into two parts, one for a discussion of each approach. In Part 1, the metarepresentation approach inspired by Shea (2014) will be discussed. As already mentioned, this approach could be used to solve the problem of triviality by appealing to Shea’s peculiar brand of metarepresentational structure in an effort to argue that noetic feelings ought to be thusly construed and so indicative of sophisticated control mechanisms. However, this approach will be demonstrated to be unfruitful due to how it fails to secure a metarepresentational construal. To make sense of Shea’s argument, I begin by introducing a barebones model from motor cognition (Section 2) and then discuss the added layers of sophistication described by the actor-critic model (Section 3). Thereafter, I discuss Carruthers’s criticism of Shea’s construal and demonstrate how this construal fails to secure a metarepresentational reading (Section 4) and conclude (Section 5).

As was introduced above, Shea’s construal of what metarepresenting consists in is non-standard. Generally, ‘metarepresentational’ describes the structure of the products of mindreading capacities: mental states (or mental representations) characterized by metabeliefs, beliefs about beliefs. In Shea’s construal, however, metarepresenting consists in cognitive systems producing subpersonal mental representations with non-conceptual content constituted by the non-conceptual content of other subpersonal mental representations.\(^\text{18}\) Going forward, reference to this

\(^{18}\) Again, ‘content’ is one of those terms in philosophy that tends to confuse because it often means something different depending on context (see Pitt 2020). Shea’s use of the term ‘content’ is different from how it is traditionally used in mindreading literature, where ‘content’ typically describes what a conscious mental state is about (e.g. in the belief that P, the belief is about P, so the content is P) and is often believed to be conceptual in nature. Though by no means non-standard, Shea’s use of the term is different, downstream from foundational studies in experimental psychology on visual cognition (see Marr 1982), in which content is not for the subject, but for the brain, or for the information processing system, used by the brain for the regulation of various cognitive processes. For example, in explaining visual cognition, the brain is said to represent an array of light across the retina, so that the content of this subpersonal brain state is the computational information pertaining to the array of light. Thus here, content is not only non-conceptual
non-standard notion of what metarepresenting consists in will be denoted by an asterisk (*) – viz. ‘metarepresentational*’ or ‘metarepresentations*’.

The purpose of Shea’s construal, however, is not to shore up metacognition against the problem of triviality, but to make plausible the case for metarepresentations* without metacognition. So, in this sense, his argument is somewhat orthogonal to our central debate. But his construal is deeply relevant for present purposes. First, his detailed defense of the metarepresentational* status of RPEs, as well as recent criticism of it by Carruthers (2020), point to how one could plausibly argue for the metarepresentational* status of any signals whose metacognitive status is up for debate, as well as, in light of these criticisms, an effective learning‐lesson about what would likely be the result of taking this approach. Second, Shea’s appeal to the prominent model of reinforcement learning, the actor–critic model, and the cognitive control mechanisms underpinning RPEs sets the stage for understanding the cognitive role approach, both its innerworkings and the problems facing it. Thus, the goal of Part 1 is twofold: expose how taking this approach will only reproduce the stalemate that already exists between proponents and opponents of metacognition, as well as set the stage for Part 2.

In Part 2, the cognitive role approach defended by Proust (2013) for shoring up metacognition against the problem of triviality will be examined. Essentially, this approach involves arguing for the metacognitive status of its candidate signals (viz. noetic feelings) by appealing to the role they satisfy in the overarching cognitive economy (and not to their status as metarepresentations or metarepresentations*). As already discussed (see Chapter 1), this role is argued to be that of guiding behavior to become epistemically sensitive, thus laying the initial foundation for properties believed to be distinct of human cognition.

The goal of Part 2 is to provide a detailed account of Proust’s cognitive role approach, specifically in an effort to diagnose the reasons behind its failure to assuage skeptics, as well as to unearth the best possible response to their criticisms. To begin, I review Proust’s theory of metacognition (Section 6) and relate it to Proust’s so-called ‘double accumulator model’ (Section 7). Thereafter, I shall demonstrate its strong family resemblance to the actor–critic model that describes the mechanisms behind reinforcement learning which are not theorized to play any

(as Shea makes clear) it is also defined solely in terms of its computational role and does not necessarily involve (or is available for) conscious report. Having said that, we now have a convenient way to summarize Carruthers’ criticism of Shea’s proposal: non-conceptual content of a computational kind with supposed meta structure does not demand a second-order construal, since a first-order construal suffices to account for it computational role, unlike the conceptual content of a personal-level kind with meta structure, whose role in cognition and behavior can only be explained by appealing to second-order states or properties. Thus, Carruthers concludes, there is nothing actually meta (second-order) about the non-conceptual content of a computational kind, or at least, not as it regards the content of RPEs (more on this below).
crucial role in facilitating either executive, reflective, or mindreading capacities (Section 8). This analysis demonstrates how Proust’s theory of metacognition tends to falls short in the eyes of skeptics, struggling to offer the sophistication needed to ward off the problem of triviality.

After demonstrating the strong family resemblance between the double accumulator model and the actor-critic model and thus demonstrating the close conceptual proximity of noetic feelings to RPEs, I shall revisit Proust’s account of noetic feelings (Section 9). It will be demonstrated that Proust’s appeal to their capacity to provide (rather than merely carry) epistemic information, interpreted to mean that noetic feelings make confidence (and its inverse, uncertainty) consciously accessible, succeeds in distinguishing noetic feelings from RPEs and thus succeeds in distinguishing the double accumulator model from the actor-critic model. Finally, I conclude with a summary of the key lessons for shoring up metacognition against criticism that will stay with us throughout the present work (Section 10).

In the next chapter (Chapter 4), I shall make the case for embodied metacognition that offers a solution to the problem of triviality and, as such, ought to put an end to the stalemate between metarepresentationalists and evaluativists. I appeal once again to recent empirical studies in affective psychology that provide evidence of an intricate brain-body mechanism responsible for the production of noetic feelings (see Chapter 2). Effectively, the sophistication of this mechanism prevents embodied metacognition from slipping down the slippery slope and entailing the ubiquity of metacognition in the animal kingdom, thus setting the stage for a plausible approach to going about securing its place as a significant precursor to self-knowledge (Chapter 6).
Part 1: The Metarepresentation Approach to Metacognition
Inspired by Shea’s Metarepresentational* Construal of RPEs

2. The Barebones Model from Motor Cognition

In the previous chapter (Chapter 2), I discussed research in cognitive and affective psychology on the nature of noetic feelings, how they are produced, how they function, etc. There, I found it helpful to introduce a taxonomy of three conceptually distinct views: the traditional view, in which noetic feelings are produced solely by brain-based mechanisms (see Reder et al, 2000), the weak embodied view, in which noetic feelings are produced by extracerebral processes and the relationship between the brain and the body is coarse-grained (see Topolinski and Strack 2009a, 2009b), and the strong embodied view, in which this brain-body relationship is fine-grained. As a reminder, I argued in favor of the strong view as it offers the most sufficient accounting of recent empirical evidence about the role of the extracerebral body in the production of noetic feelings (see Allen et al. 2018; Chua and Bliss-Moreau 2016; Fiacconi 2016, 2017). In what follows, this taxonomy will be applied to the debate over the metacognitive status of candidate signals.

Before it can be discussed how to fit RPEs into the metamnemonic framework articulated in the previous chapters (Chapters 1 and 2), I remind that this discussion will involve a close reading of Shea (2014), and though he does not intend RPEs to be conceived of as metacognitive, his approach to arguing for the metarepresentational* status of RPEs can serve as a model for how one might argue for the metarepresentational* status of noetic feelings in an attempt to defend metacognition against the problem of triviality. This is because, as skeptics are happy to concede (e.g. Carruthers 2009b), if noetic feelings were indeed metarepresentational or if the cognitive states characterized by noetic feelings approximated some metarepresentational structure, then metacognition would be sufficiently robust to ward off the problem of triviality. So, the metarepresentation approach is very much active in this debate.

Moving on to discuss details surrounding RPEs, depending on the theory of their implementation, RPEs can satisfy several distinct roles at differing levels of description. The theory that perhaps entails the least conceptual commitments is the one that construes RPEs as the products of cognitive control mechanisms, not unlike those which form the basis of Proust’s philosophy of mind (see Chapter 1). Meanwhile, more sophisticated theories describe RPEs also
at the neurobiological level of description, as produced by specific neuroanatomic structures and propagated by specific neurotransmitters (see Schultz 1998). But what concern us here is how best to understand RPEs from the perspective of cognitive psychology, so our discussion will focus primarily on how RPEs are implemented in cognitive control mechanisms, in particular how RPEs are described by the prominent actor-critic model (Dayan & Abbot 2001; Sugrue 2005).

To illustrate the technical details behind the actor-critic model, I shall steadily introduce more layers of sophistication, starting with a barebones model based on Conant and Ashby’s (1970) seminal work in the cybernetics of motor cognition. But bear in mind that this model will become modified to describe more complex actions (see Section 3). Let us begin by describing the act of pushing a coffee mug across a desk, such that there is the bodily action $\varphi$ with the goal state $G$ (i.e. the end result of the coffee mug moved to its intended location).

According to any cybernetic story, before $\varphi$ is executed, the cognitive subsystem responsible for monitoring and controlling $\varphi$ will deploy two models: (i) an inverse model that transforms the expected sensory consequences $f$ (e.g. the proprioceptive information required to produce the movement that will place the coffee mug at its intended location) into the motor command $\varphi$ that will be deployed to achieve $G$, and (ii) a forward model that describes the transformation from $\varphi$ back to $f$ that enables the prediction of the sensory consequences estimated to occur during the performance of $\varphi$.

While $\varphi$ is executed, the expected sensory consequences $f$ will be compared to the observed sensory consequences $f'$ (i.e. the sensory consequences that actually occur during the performance of $\varphi$). In the event of a discrepancy (say, the coffee mug is heavier than anticipated and more force is required, such that $f' \neq f$), a feedback signal $S$ will be produced that updates both models: (i) the inverse model to ensure the obtaining of $G$ and (ii) the forward model to ensure better predictions about the sensory consequences of $\varphi$ going forward.

As this example makes clear, $S$ could serve as a candidate signal for metacognition, since, to some degree, $S$ can be said to be about this cognitive control process (viz. cognition about cognition: metacognition). But to base metacognition on a mechanism as simple as this would clearly invite the problem of triviality and thus divorce metacognition from any crucial role in producing cognitively sophisticated behavior, not to mention robbing metacognition of any particular significance in explaining the origin of self-knowledge.

In the next section, layers of sophistication will be added to this barebones model to transform it into the actor-critic model with the goal of illustrating how cognitive control
mechanisms are argued to underpin capacities for more sophisticated forms of cognition and behavior, capacities that begin to resemble capacities for metacognition and mindreading. In Part 2, I shall compare the actor-critic model to Proust’s (2013) double accumulator model for describing mechanisms for metacognition. There the goal will be to diagnose the reasons why her cognitive role approach for shoring up metacognition is often met with criticism.

3. The Sophistication of the Actor-Critic Model

As already introduced, Shea (2014) appeals to the actor-critic model to argue for the metarepresentational* status of RPEs; meanwhile, a similar appeal could be made to defend metacognition against the problem of triviality. Effectively, one might argue that the candidate signals of metacognition, noetic feelings, are metarepresentational*, suggesting noetic feelings characterize cognitive states that approximate metarepresentational structure, from which point one might argue that this structure is thus sufficiently sophisticated to solve the problem of triviality. In what follows, the reasons for why this appeal to metarepresentation would fail to convince skeptics shall be diagnosed and, in so doing, the stage will be set for diagnosing the reasons for why Proust’s cognitive role approach continues to encounter criticism.

To be clear, Proust (2007) argues that the actor-critic model is insufficient to explain the cognitive control structures described by her theory of metacognition. But, as I shall demonstrate in Part 2, there are nonetheless remarkable similarities between the two models and bringing this family resemblance to light will illuminate criticisms facing her theory of metacognition. What is more, since the present work’s case for embodied metacognition takes Proust’s theory as a starting point, diagnosing the reasons behind why her theory is often met with criticism will assist the achievement of the present work’s goal of offering a solution to the problems facing metacognition.

Moving on to explain the actor-critic model, as the name suggests the actor-critic model consists of two distinct cognitive subsystems, the actor and the critic. These systems are appealed to describe and explain the cognitive process behind reinforcement learning, the reward/punishment of action selection. The combined cognitive architecture underpinning reinforcement learning has the objective "to learn to act so as to produce sensory input for which the primary reinforcement value is maximized" (Barto 1995). Thus, the key to understanding this process involves determining how the primary reinforcement value is computed, and this is done by contrasting the performance of the actor to the control objectives of the critic.
Consider again the barebones model from motor cognition. In the actor-critic model, the cognitive subsystem responsible for monitoring and controlling $\varphi$ is referred to as ‘the actor’. This subsystem (same as above) deploys an inverse model describing how to execute $\varphi$ so as to achieve $G$, as well as a forward model that enables the prediction of the sensory consequences of $\varphi$ when $G$ is the aim. Recall that (same as above) in the process of performing $\varphi$, discrepancies between the predicted sensory consequences and observed sensory consequences will generate a feedback signal $S$ for revising both models to ensure better performance and better predictions.

A layer of sophistication is added to this story by the introduction of a distinct cognitive subsystem known as ‘the critic’. Its role is to monitor the actor’s performance of $\varphi$ and produce a second layer of feedback $\delta$ known as ‘reinforcement feedback’. This feedback is ultimately the evaluation of action-selection according to the critic’s own specific control objectives relative to that action: $(C^\varphi_1, ..., C^\varphi_J)$. Thus, the crucial difference between the barebones model and the more sophisticated actor-critic model are twofold: the introduction of the critic subsystem (with its control objectives) and a novel feedback signal, namely $\delta$.

Critical for our discussion, $\delta$ is the signal that Shea (2014) argues ought to be construed as metarepresentational*. But keep in mind the caveat above that while Shea does not argue that $\delta$ is a metacognitive signal, one could adopt his strategy and argue for the metarepresentational* status of noetic feelings in an effort to solve the problem of triviality facing metacognition. But to properly diagnose the reasons for why this approach will likely fail to convince skeptics (see Carruthers 2020), the technical details behind Shea’s argument will need to be thoroughly introduced.

Let’s begin by making sense of the added layer of sophistication behind the actor-critic model. Imagine an actor portraying Hamlet in the famous “To be, or not to be” soliloquy in Act 3, Scene 1 (Shakespeare 1599-1601/2008). The actor will undoubtable have some personal measure for how well he is executing the performance (honied by years of practice and feedback), which he will use to gage his performance in real time. Say, something happens to be off about this particular performance, say, his heart just isn’t in it for some reason (perhaps, this is his third performance of the night and he’s exhausted or bored). Assuredly, he will become conscious of this slack and, being a good thespian, will no doubt adjust his performance by, for example, enunciating with more gravitas. So far the feedback process only describes the initial barebones model from motor cognition, so let’s now introduce the added layer of sophistication provided by the critic subsystem.

Imagine a renowned theater critic is sitting in the audience observing the actor’s performance. Her job it is to offer an evaluation $\delta$ of the actor’s performance according to two
objectives (there could be more objectives than two, but two should suffice for present purposes). First, her criticism will concern how well the actor is executing the soliloquy. This captures what is referred to as the ‘tactical objective’ and reinforcement signals instrumental in meeting the tactical objective will only pertain to the immediate goal G, that of giving a good performance of the “To be, or not to be” soliloquy.

Secondly, the critic will also need to factor in how well the actor is portraying Hamlet, which captures, what is often called, ‘the strategic objective’. Here, reinforcement signals will not only pertain to the immediate goal G but also to the long-term goal H (i.e. that of giving a good portrayal of Hamlet). In order to determine whether the actor’s performance has reached the strategic objective, the critic will consider a base standard, one that applies to any scene featuring Hamlet, informed by previous portrayals of Hamlet, as well as considerations about future portrayals. But for the sake of brevity, an explanation of the technical details behind this aspect of reinforcement learning will be omitted.

For the analogy to cohere with the actor-critic model, a few things will need to be stipulated. First, our focus is on the actor’s performance of Hamlet before and after learning of the critic’s feedback. As to that, let us stipulate that the feedback is mostly positive with the exception of the nunnery scene “feeling uninspired”. Second, we will need to imagine the critic’s feedback δ as delivered in real time, as though she were able to whisper surreptitiously into the actor’s ear during the performance. Third, the critic will need to have direct access to feedback S, which the actor receives, as though she were wired up to same cerebral circuits and receiving a readout of how well the performance is unfolding according to the actor’s own internal standards. Let us stipulate further that upon hearing the critic’s feedback δ, the actor is receptive to it, causing him to discover within himself a pathos that will infuse the soliloquy with the missing vitality.

The crucial feature illuminating the mechanisms behind reinforcement learning is thus δ, in particular how δ will not only shape the actor’s selection of subsequent actions, but how it will also shape the actor’s own internal standards for performance. How best to make sense of this fact about the actor-critic model as it pertains to the best description of δ at the cognitive level (in particular, how best to construe the content of δ so that it can shape the actor’s own internal standards) is at the core of Carruthers’s criticism of Shea’s argument. So this point about how best to account for the role of δ in reinforcement learning will be discussed in more detail below.

Finally, the critic’s overall goal is to provide reinforcement feedback δ that ensures that the actor, in meeting the demands set by the immediate performance of the soliloquy, viz. the tactical objective, actually learns to give the best overall performance of Hamlet, viz. meets the strategic
Chapter 3: Problems Facing the Defense of Evaluative Metacognition

This is essentially what is meant by "to learn to act so as to produce sensory input for which the primary reinforcement value is maximized" – the primary reinforcement value is effectively the product of meeting the strategic objective (i.e. the best performance, in the long run).

Now, in order to arrive at an understanding of the crucial nuances behind the actor-critic model, let us add the layers of sophistication to the barebones model from motor cognition above in order to capture the details offered by this illustration. Recall that nothing about the initial story has changed: in executing \( \phi \) to achieve \( G \) (pushing the coffee mug across the table), the control subsystem, viz. the actor subsystem, determines the sensory consequences \( f \) and receives a feedback signal \( S \) to update its models to produce and predict the next action in the sequence \( \phi' \).

As already mentioned, the added layer of sophistication behind the actor-critic model come from the introduction of the critic subsystem and its reinforcement signal. Meanwhile, the role of the critic is to reinforce the selection of \( \phi \) and shape subsequent actions in the sequence (\( \phi' \ldots \)) by, effectively, changing the value assigned to \( \phi \). In other words, actions are assigned a value based on two levels of feedback: the initial feedback signal \( S \) and the reinforcement signal \( \delta \), and the crucial layer of sophistication is thus accounted for by how \( \delta \) determines this value.

The initial feedback value assigned to any action is specified by the discrepancy between what is observed and what is expected. In our example, this is captured by the difference between \( f' \) and \( f \), so that \( S = (f' - f) \). Crucially, the critic can influence the initial feedback value in one of two ways, but always with respect to the critic’s control objectives, \( C_i^{\phi}, \ldots, C_j^{\phi} \). The value assigned to the initial feedback signal will be increased if the dynamics of the action correlate with descriptions found in \( C_i^{\phi}, \ldots, C_j^{\phi} \) and the value assigned to feedback will be decreased if there are discrepancies between the observed dynamics of the action and those dynamics described by the control objective. This is thus the essence of reinforcement learning (see Sutton and Barto 1998).

The critic subsystem receives the same information that the actor does about how executing \( \phi \) calls for \( f \) and the same feedback signal \( S \) concerning the effectiveness of \( f \) in achieving \( G \). But the critic subsystem compares this information to its control objectives: \( C_1^{\phi} \), the tactical objective of achieving \( G \), and \( C_2^{\phi} \), the strategic objective of achieving the best overall performance, \( H \) (in our example from motor cognition, \( H \) is the set of actions involving moving objects of similar weights in a similar manner across surfaces with similar resistance). The outcome of this comparison is a reinforcement signal that either “rewards or punishes” the actor for its selection of \( \phi \) with \( f \) to achieve \( G \). Should the selection of \( \phi \) with \( f \) correlate with the information described by the critic’s
control objectives, then its selection of \( \varphi \) will become positively reinforced and the upshot will be a reinforcement signal \( \delta \) that is said to reward the selection of \( \varphi \).

In the best-case scenario, \( \delta \) will ensure the actor obtains the strategic objective \( H \). Crucially, this is made possible due to how \( \delta \) is used by the actor to adjust the value of \( \varphi \), and this revised value is then subsequently used for the selection of the next action in the sequence \( \varphi' \):

\[
\varphi \rightarrow \varphi' = \varphi + \delta = \varphi + (S - (C_i^\varphi, \ldots, C_j^\varphi))
\] (1)

In other words, by comparing the actor’s performance to its own control objectives, the critic is able to provide feedback that not only concerns how well the immediate performance is faring, but also how the performance fares with respect to the outcome of previous performances, as well as predicted outcomes of future performances. Now with that explanation on the table, \( \delta \) is robust enough to meet the conditions of an RPE signal for present purposes. To be clear, the feedback signal \( \delta \) in (1) will be compared in Part 2 to the feedback signals (viz. noetic feelings) described by Proust’s double accumulator model (see (2)).

4. Controversy Around a Metarepresentational* Construal of RPEs

Let us return to the problems facing the defense of metacognition, in particular a defense based on an appeal to the metacognitive status of its candidate signals. To that end, the controversy surrounding the construal of RPEs as metarepresentational* will be discussed below. Though Shea (2014) makes a strong case for this construal, his approach is criticized by Carruthers (2020) as failing to demonstrate how RPEs are anything over and above mere first-order cognitive signals and so not sufficiently ‘meta-’ to be properly construed as metarepresentational* (let alone, metarepresentational in the mindreading sense). To understand this disagreement, and how it applies to our discussion about how to defend metacognition against criticism, the details of Shea’s argument will be analyzed below with help of our explanation of the actor-critic model above.

As it turns out, Carruthers (2009b; 2012; 2017) employs the same reductive strategy to argue against Proust’s theory of metacognition. Ultimately, Carruthers argues that in order to account for their role in cognition and behavior, only those mental states (or mental representations) that are produced by mindreading capacities demand a second-order construal, as
representations about representations (beliefs about beliefs). Meanwhile, since mental states characterized by noetic feelings are not the products of mindreading capacities, an explanation of their role in cognition and behavior, Carruthers argues, does not demand a second-order construal, and thus these mental states are not cognitive states about other cognitive states and so not metacognitive states (see Carruthers 2017). As such, Carruthers’ reductive strategy has the potential to concern the case for embodied metacognition as well, since I shall seek to make it plausible that some first-order cognitive states ought to be construed as metacognitive, despite not being cognitive states that are about other cognitive states in this mindreading sense. So, after Shea’s argument is analyzed, we shall discuss the details surrounding Carruthers’ criticism of it.

Let’s begin by making sense of Shea’s peculiar understanding of metarepresentations*. RPEs, he argues, are metarepresentational* while also being non-conceptual. Conceptual representations are constituted by concepts and have propositional content determined by truth conditions, content that is described by whatever follows the that-clause in believes that p (see Pitt 2020; also see our discussion above in Chapter 1). But non-conceptual representations do not have semantically-significant constituent structure and so are not constituted by concepts and do not have propositional content determined by truth conditions (see Evans 1982; Peacock 1992; Bermúdez 2020). Nonetheless, non-conceptual representations do not exclude representational properties determined by conditions for correctness.

As such, Shea suggests two distinct accounts for construing a non-conceptual representation M as metarepresentational*, i.e. as representing another representation R. However, he argues that only one account actually justifies construing M as a metarepresentation*. This is because, in order to be a proper metaterepresentation it must be “a representation whose content concerns the content of another representation” (ibid 315; see Proust 2013 and Carruthers 2017).19

According to the first account, M could represent the vehicle properties of R. This would mean that the correctness conditions for M would be specified by causal, non-semantic properties of (in this case) neural structures that the relevant cognitive mechanisms are sensitive to (Shea 2007; 2018; Drayson 2018). Examples of such properties are the relative firing strength of neural networks or the variance in firing rate of neural assemblies, both of which are clearly non-semantic, causal properties of neural structures (see de re reference discussed in the Introduction).

---

19To reference a comment made in an earlier footnote, there is something approaching an equivocation here between content, in the sense of personal level states, and content, in the sense of brain states. Typically, when metarepresentations are discussed, one deploys the former sort of notion, but Shea does not discuss personal level states at all (RPEs are not theorized to be accessible for conscious report), so he must be employing the subpersonal (and non-conceptual) notion of content, and hence the “switch” from metarepresentations to metarepresentations*.
However, Shea argues that this representational structure is insufficient to warrant a metarepresentational* construal of $M$ since it does not involve any representational property of $R$ as factoring into the correctness conditions for $M$. To be properly metarepresentational*, the correctness conditions for $M$ will need to concern the content of $R$ (and not merely its vehicle properties), so that $M$ can be construed as representing $R$ as a representation (see de dicto reference discussed in the Introduction).

The example that Shea gives describes the content of $M$ as “the current visual representation of the location of the light is likely to be false” (ibid: 316). Thus construed, the content of $M$ clearly concerns a representational property of $R$ as it explicitly describes it is as a representation, i.e. capable of being misrepresented. But crucially, this content also includes the content of $R$, namely the location of the light, which, as the case may be, is either correct or incorrect, veridical or illusory, and thus reflects a semantic, rather than causal, property of $R$.

Shea’s argument for this metarepresentational* construal has two prongs. One prong amounts to appealing to a variant of scientific realism, effectively pointing out how prominent theories across the brain sciences depend upon an appeal, for their explanatory power, to descriptions of RPEs as metarepresentational*. Though this argument might seem compelling in its own right, it does not concern our present discussion that focuses on how best to ward off Carruthers’ reductive strategy, nor does it apply to noetic feelings of metacognition which are not supported by scientific theories in this manner (i.e. theories about noetic feelings do not depend on their construal as metarepresentational* for their explanatory power).

The second major prong of Shea’s argument actually aims to anticipate Carruthers’s criticism. In drawing upon Carruthers’s (2009b) criticisms of Proust’s strategy (see below), Shea attempts to nip these attacks in the bud by appealing to a Dretske-style test (1988), which Shea construes as, from the outset, meeting Carruthers’ own criteria for how best to think about metarepresenting in the usual mindreading sense, entailing de dicto reference.

By arguing that RPEs pass the Dretske-style test, Shea argues for how best to specify the content of RPEs (by way of arguing for how best to specify their correctness conditions), while also arguing that RPEs are metarepresentational*. RPEs represent that “the reward [is] more/less than the represented value” (ibid: 332). Thus, Shea defends this construal as qualifying RPEs as metarepresentational* due to how their content concerns the content of another representation, namely the represented value of the selection action provided by the initial feedback signal.
For Dretske, representational content is fixed by determining how representations are used to guide the learning process, specifically the learning of adaptive behaviors (see Dretske 1988, 2006). So, the test amounts to determining the specific conditions under which the representation in question serves to reinforce behavior, i.e. rewards or punishes action selection. RPEs are only beneficial to the learning process if reinforcement turns out to be more/less than the value described by the initial feedback signal, otherwise RPEs cannot serve to reinforce action selection.

Let’s revisit our illustration above. The critic’s feedback about the actor’s performance needs to both concern and supersede the actor’s own internal feedback, otherwise the critic’s feedback will fail to reinforce, and so control, the actor’s performance going forward. This relationship between RPEs and the initial layer of feedback thus explains the role RPEs play in reinforcement learning. For this reason, Shea argues, following Dretske’s framework for how best to determine representational content, this is how the content of RPEs ought to be construed: RPEs are correct (and, insofar, only useful to the cognitive system) if they tend to represent the initial feedback as either better than or worse than how it was initially represented: “the reward [is] more/less than the represented value”.

Recall that the reinforcement signal $\delta$ sent by the critic influences the initial feedback value. Recall $\delta$ influences this value by either increasing or decreasing the initial value depending on the degree to which the feedback produced by the selected action correlates with the critic’s own control objectives (see (1) above). In other words, the critic sends $\delta$ to the actor to reinforce its action selection, so that the actor learns, in the long run, how to conform its behavior to the control objectives, effectively “maximiz[ing] a weighted sum of all future primary reinforcement values” (Barto 1995: 7). So it follows that $\delta$ is used by the cognitive system as a guide for the learning, and so a Dretskeian ought to construe RPEs, at the very least, as representational, as having content of the non-conceptual, subpersonal variety.

The controversy surrounds how best to construe the content of $\delta$, that is, how best to construe the exact message the critic communicates to the actor in the process of reinforcing action selection. Is the best construal of the content of $\delta$ “the reward is more/less than the initial value”? If so, $\delta$ is metarepresentational* due to how it represents the content of the initial feedback signal. But perhaps this construal is a hyper intellectualization of the content, and a much less sophisticated construal will suffice to account for $\delta$’s role in the learning process. Carruthers (2021) argues that this role can be explained equally well by a thinner construal of the content: in order to reinforce action selection, the critic need only communicate “increase/decrease reward by [x]”.

Consider once again the illustration above. Carruthers’ reductive strategy amounts to arguing that the critic need only say something along the lines of “Good!” or “Bad!” (with more or less gumption) while the actor performs the soliloquy in order to reinforce action selection. Thus, the critic need not comment on the adequacy of the actor’s own internal standards, as though saying, “your own internal standards for this performance are good” (or bad, as the case may be).

Of course, this thinner construal raises questions about how \( \delta \) can alter the initial value of the feedback if it fails to represent it. But if the actor subsystem is configured to receive \( \delta \) as input and alter internal standards accordingly, yielding altered values going forward, then \( \delta \)’s role in guiding the learning process can be explained by appealing to how it interfaces with specific configurations of the actor subsystem rather than through the second-order content of \( \delta \). If this is the correct construal of the content of \( \delta \), then RPEs have a content that is, at their most sophisticated, about the vehicle properties of the initial feedback, but not about its semantic properties, and so, by Shea’s own standards (see above), RPEs possess a content of the first-order, rendering RPEs non-metarepresentational* (and so certainly not metarepresentational).

The purpose of this discussion is not to evaluate how best to construe the content of these representations, though this is admittedly already baked into the narrative somewhat through the prior distinction between metarepresentations* and metarepresentations. Rather, the goal is to offer a diagnosis about what plagues the metacognition debate and, in so doing, make inroads into developing a possible antidote. In this regard, what stands out about this controversy is that though Shea’s appeal to the Dretske-style test was an explicit attempt to circumvent Carruthers’ reductive strategy, it ultimately made him vulnerable to it in the end. So what went wrong?

Simply speaking, the appeal to the Dretske-style test is unsuccessful in securing a metarepresentational* reading of the content of \( \delta \) because it failed to secure a metarepresentational (sans asterisk) reading of the content. In particular, the actor-critic model fails to describe a cognitive process with the required sophistication to demand a metarepresentational construal of its reinforcement signals. Specifically, the role of the reinforcement signals in the overarching cognitive process can be explained without appealing to their content as possessing second-order representational structure, as possessing content about content.

Consider how different it would be if the actor and critic were distinct persons that often depended on capacities for mindreading for communication. In this case, little doubt would surround the proposal that some instances of communication demand a metarepresentational construal, representing representations as representations, and thus implying a metarepresentational* construal, content about content. For example, if the critic’s initial feedback
failed to bring about the desired outcome, the actor might refer back to the feedback as the previous feedback it was: “When you said [x], what exactly did you mean?”. Such acts of communication, namely the clarification of misunderstandings, clearly demand a metarepresentational reading as they clearly demand capacities for mindreading.

Though the actor and the critic are distinct levels of the control structure, they are still wired together and capable of “speaking on the same level”, capable of referring to the properties of shared commodities without needing to recast messages in the second-order, and so the subsystems do not need to represent certain properties as the properties they are in order to ensure messages are received as intended. Carruthers thus exposes how the de re mode of reference suffices to describe the communication between the actor and critic subsystems, and the appeal to the more sophisticated de dicto mode of reference, in which properties are represented as what they represent, is a hyper-intellectualization of the communication process.

Thus, what is needed to ensure this process receives a metarepresentational construal are reasons to believe communication between the two subsystems is inevitably susceptible to miscommunication, and, furthermore, reasons to believe the two subsystems not only have the tools to mitigate miscommunication, but also actively utilize these tools in the process. In other words, missing from this picture is the need to appeal to the sophistication of concepts (or concept-like entities) in explaining the communication between subsystems. On this point, Carruthers agrees: “representation as […] requires concept-like states that can be components of belief-like ones” (Carruthers 2020: 2340).

Appealing to concepts might represent a possible avenue for arguing that certain cognitive signals are metarepresentational* but, regarding the debate over the status of metacognition, such an appeal cannot work. Conceptual capacities are too sophisticated of a demand on the representational systems in question if they are to be operative in the minds of animals and serve as precursors to self-knowledge. This is because the emergence of conceptual capacities are indeed part of what one intends to explain by appealing to metacognition in the first place (see Proust 2013). So, this proposal goes beyond the bounds of what ought to be sought after when defending evaluative metacognition. But by pointing to concept-like states and belief-like states, Carruthers allows for some conceptual room for a precursor to metarepresentational metacognition. This suggests an end to the stalemate between evaluativists and metarepresentationalists might involve appealing to cognitive control structures that regulate states characterized by proto self-beliefs. Indeed, this the approach advocated by the present work below (see Chapter 4).
5. Conclusion to Part 1

As the discussion above and the discussion in Chapter 1 make clear, both RPEs and noetic feelings are conceived of as products of cognitive control mechanisms of modest sophistication, feedback signals that guide adaptive behavior and whose metacognitive status could be up for debate as a result. However, the mechanism behind RPEs is clearly not sufficiently sophisticated so as to play a crucial role in enabling and facilitating an executive or reflective process, and so RPEs cannot possibly serve as candidate signals for metacognition. This suggests that in order for noetic feelings to serve as candidate signals, it will need to be demonstrated how their underlying mechanism facilities and enables executive control over cognition and behavior.

Before concluding this part, I wish to point out how the actor-critic model is akin to models that motivate traditional views discussed in the previous chapter (Chapter 2). This is because of how both kinds of model describe signals produced solely by brain-based cognitive control mechanisms. This is not to say that Shea’s approach concerns noetic feelings. It certainly does not. But by exposing a family resemblance between models of RPEs and models behind the traditional view of noetic feelings, light is cast on the path out of the quagmire, namely a path that involves making a hard break with the traditional view of noetic feelings and instead appealing to them as strongly embodied, conscious expressions of emotion.

Herein lies the crucial difference distinguishing noetic feelings from other feedback signals, a point overlooked by traditional views. Noetic feelings are feelings, and so conscious expressions of emotion, whose role in cognition and behavior cannot be fully accounted for without appealing to extracerebral body (see Chapter 2). However, this is not how noetic feelings are theorized to work according to the traditional view, which appeals solely to cerebral mechanisms to account for how noetic feelings are produced and integrated into the cognitive process. Neglecting their embodied status means failing to account for the cognitive sophistication within their underlying mechanism, as well as their crucial role in the emergence of self-knowledge.

In the next part, I argue that according to Proust’s account of noetic feelings, the relationship between cerebral and extracerebral processes can be characterized as very coarse-grained, making it potentially weaker than some weak views discussed above (see Chapter 2). This is because her theory attempts to account for the role of noetic feelings in cognition and behavior without appealing to their status as crucially intertwined with extracerebral processes. For this reason, I refer to Proust’s construal of noetic feelings as ‘weakly embodied noetic feelings’.
As we will see below, Proust takes a different approach to arguing for the metacognitive status of noetic feelings than the approach that one might take if Shea’s argument for the metarepresentational* status of RPEs were pursued. Proust does not appeal to the content of noetic feelings as metarepresentational* (or metarepresentational), but rather appeals to their role in enabling a normativity of an epistemic kind, and thus making possible certain distinctly-human properties of cognition (see Chapter 1). But as will be discussed in detail below, this approach encounters its share of criticisms (Carruthers 2009, 2017, 2020).

Part 2: The Cognitive Role Approach and the Status of Weakly Embodied Noetic Feelings

6. Revisiting Proust’s Theory of Metacognition

In this part, I shall demonstrate how Proust’s theory of metacognition encounters criticism by exposing the close conceptual proximity between her account of noetic feelings and the standard account of RPEs, in particular by showing the conceptual similarity between the actor-critic model discussed above and Proust’s double accumulator model, which she employs to describe the role of noetic feelings in cognition and behavior. For now though, it will help to reintroduce her theory of metacognition, as well the role she envisions noetic feelings to play.

Recall from Chapter 1 that Proust’s approach to understanding the nature of metacognition is downstream from the cognitive psychology of metamemory (see Koriat 1993, 1995, 2000). Here, metacognition is cognition about cognition in the specific sense that it involves dedicated cognitive mechanisms that monitor and control cognitive processes. Both monitoring and control functions are made possible through the production of evaluative signals, which, according to Proust, can potentially be expressed through dedicated somatic markers, either conscious or unconscious patterns of bodily activity (see Damasio 1994). Thus according to Proust, noetic feelings as the products of cognitive control mechanisms that have the function of either evaluating the outcome of overt cognitive acts and their underlying covert processes or predicting their potential success. Examples of these acts are memorizing a list of words, recalling a name from memory, discriminating between two visual stimuli, etc.
Thus, noetic feelings are for Proust (and others as well, e.g. Koriat 2000) non-conceptual representations whose correctness conditions are determined by specific vehicle properties of neuronal activity. Crucially, noetic feelings represent in a distinct sense from how propositional representations are often said to represent. This is because noetic feelings do not refer in the proper sense, as their content does not possess semantically-significant constituent structure. Rather, their content is constituted by certain non-semantic, vehicle properties of neuronal activity, such as the strength of a memory trace (see Proust 2003; Koriat 1993). Recall that it was thus concluded above (see Chapters 1 and 2) that the representational structure of noetic feelings ought to be construed as one of indexing vehicle properties of neuronal activity (see de re instead of de dicto mode of reference; Strawson 1959; Quine 1956; Burge 1977; Proust 2013).

Recall also that this non-conceptual representation was discussed above as a cognitive affordance (specifically in Chapter 1), which, according to Proust, comprises a cognitively light species of mental representation that presents, at the animal-level of description, a functional relationship between an internal capacity (e.g. the capacity to recall an item from memory) and internal features produced by the cognitive system itself (e.g. patterns of mnemonic activity). All in all, cognitive affordances and noetic feelings are different sides of the same coin, as the content of both is structured by correctness conditions determined by vehicle properties of neuronal assemblies (e.g. firing rate of neuronal populations), which can be expressed by imperatives, such as “Recall!” or “Don’t Recall!”.

7. The Double Accumulator Model

Let us turn to Proust’s double accumulator model. As the name suggests, this model describes two crucial subsystems: the sensory and the control accumulators or comparators (also called the ‘primary’ and the ‘secondary’ accumulators, respectively), which she envisions as comprising the enabling conditions for metacognitive governance, which, in turn, is the lynch pin to her argument for why metacognition ought to be construed as a significant precursor to self-knowledge (see Chapter 1). Though Proust’s theory of metacognition was discussed in detail above, here the focus will be on its strong family resemblance to the actor-critic model. Exposing this similarity will demonstrate how her theory of metacognition is subject to the problem of triviality.
Proust proposes the double accumulator model to describe the mechanism that produces noetic feelings (see Proust 2013: 99-109; for a similar reading of Proust’s theory, see Langland-Hassan 2014). The sensory comparator has the function of making decisions for the current task. It does this by computing the so-called ‘balance of evidence’ through issuing predictions and receiving feedback aimed at correcting those predictions (see Stanislaw and Todorov 1999). Meanwhile, the control comparator has the function of extracting information from successive trials and using this information to calibrate predictions issued by the sensory comparator.

Proust tells us that, whereas the sensory comparator enables the making of “decision[s] under uncertainty”, it is the comparison performed by the control comparator that enables the “assessing of one’s uncertainty” (ibid: 104). Crucially, it does this by producing non-conceptual, non-metarepresentational representations, signals, about the probability of error/accuracy of the relevant cognitive task, which enables the calibration of cognitive performance to the parameters set by the control comparator (ibidem). Calibrating cognitive performance (of both overt cognitive actions and covert cognitive processes) to parameters set by the control comparator makes possible metacognitive governance (see Chapter 1).

But before metacognitive governance is reintroduced, take note of the conceptual similarity between this signal described above and the reinforcement signal described by the actor-critic model (see (1)). Both are non-conceptual, non-metarepresentational\(^{20}\) representations that inform the cognitive system about the error/accuracy of cognitive operations, used to calibrate performance by ensuring the system meet criteria set by a control mechanism. So far, the only difference is that of a change in label, from critic to control comparator, which means any real difference between the models must lie in the parameters set by the control comparator, specifically those that make metacognitive governance possible.

Thus, to grasp the details of the double accumulator model, the parameters set by the control comparator will need to be thoroughly discussed. Proust (2013; see 99-109) introduces three dimensions of context-sensitivity demanded by metacognition. To be clear, her explananda are those advanced decision-making behaviors observed under the uncertainty monitoring paradigm (see Introduction). In order to achieve this level of sophistication, the cognitive system must possess certain properties that enable the cognitive flexibility demanded by such behaviors.

First, success in the uncertainty monitoring paradigm demands instrumental context-sensitivity, which, in turn, demands the cognitive system detect new regularities and update its

---

\(^{20}\) Take note that this is non-metarepresentational without the asterisk. That said, Proust also does not think noetic feelings are metarepresentational\(^{1}\), since she argues that they do not have the content of another representation as their content, however “content” is construed (personal/subpersonal or computational role/aiming at correctness).
internal models for the selection of adaptive behavior. Second, the cognitive system must also feature strategic context-sensitivity, in which the system distinguishes signals from noise in various circumstances and performs a costs-benefits analysis about whether to act, and if so, how best to respond. Both of these forms of flexibility result from a cognitive system that possesses solely the sensory comparator (and not the control comparator).

Instrumental and strategic context-sensitivity pave the way for the third and final type of context sensitivity, which Proust simply refers to as ‘metacognitive context-sensitivity’. Metacognitive context-sensitivity, in turn, depends upon two properties of the cognitive system. First, the cognitive system needs to possess the capacity to produce and make accessible a subjective likelihood that a specific cognitive operation was successful or unsuccessful (needs to be able to make both accessible). Here, Proust can be read as equating this subjective likelihood with the non-conceptual, non-metarepresentational signal above (i.e. the candidate metacognitive signal). Moreover, it is likely she conceives of this subjective likelihood as a noetic feeling. This is because, subjective likelihood is formatted so as to be accessible for the purpose of action selection, possibly even accessible at the animal-level of description in the form of a conscious feeling.

That said, it is an open question whether this signal expressing subjective likelihood is equivalent to a noetic feeling, since it is unclear whether this signal, as Proust describes it, is conscious or unconscious, or whether it incorporates somatic markers.21 One crucial aspect about this signal is that it stands in contrast to objective likelihood, in that it could potentially misrepresent the actual (i.e. objective) likelihood of cognitive success. This suggests, the signal is not subjective in the sense of being accessible to the subject, but rather subjective in the sense of not necessarily correct, and so able to misrepresent, and thus reflects some species of mental representation. But, of course, not all representations are represented at the animal-level of description, so it is deeply unclear how best to square this signal with noetic feelings. Thankfully though, the argument of the present work does not depend on resolving this issue (see Chapter 4).

Second, metacognitive context-sensitivity also concerns the capacity of the cognitive system to select an appropriate epistemic norm for the particular learning context, the selection of which, itself, can be done in one of two ways (for more on epistemic normativity within Proust’s theory, see Chapter 1). On the one hand, the control accumulator can prescribe a norm based on the action being called for. For example, should the task be to reconstruct a forgotten shopping

---

21 I have argued above and shall argue below that noetic feelings ought to be construed as strongly embodied, incorporating conscious somatic markers as conscious expressions of emotions, and that this construal will assist us in solving the problem of triviality and the ultimate problem of the origin of self-knowledge. So, I do not mean to be uncharitable. It is rather that, as I read Proust, she does not commit to construing them as such and, moreover, does not articulate the advantages this construal brings.
list, its success will be determined by the degree to which a full list can be reconstructed. As such, the control accumulator will deploy the epistemic norm of exhaustiveness. But if the task is instead to determine the exact amount of money owed, then the norm of accuracy will be deployed.

The second way that the epistemic norm can be selected is during the execution of the cognitive operation. For example, imagine calculating a sum in your head only to realize that if you truly want the exact amount, you will need to do the sum on paper, so you decide, in the process of calculating, that a mere estimate will suffice. Notice this scenario illustrates a change in epistemic norm during the execution of a cognitive operation: you went from wanting to have the exact sum (accuracy) to wanting to be able to do the sum quickly and easily (fluency). That said, one might argue fluency is not an epistemic norm but rather an instrumental one, but as Proust (2013) makes clear (see 129-130), fluency should be construed as an epistemic norm due to how it approximates accuracy: fluent behaviors are fluent in virtue of their tendency to be sufficiently accurate.

In order to explain the change in epistemic norms, one can appeal to a change in confidence. In the above example, this was the change in confidence with respect to one’s ability to perform the sum in the head. This brings us back to the first property of metacognitive context-sensitivity. Within Proust’s theory of metacognition, confidence is mediated by a candidate metacognitive signal that describes the subjective likelihood of the success of the cognitive action. Thus, due to its potential to enable highly flexible cognitive action by the on-the-fly adjustment of epistemic norms through the production of feedback signals, which describe the subjective likelihood of success, metacognitive context-sensitivity is described as enabling metacognitive governance, the crown atop Proust’s theory of metacognition.

8. The Double Accumulator and the Actor-Critic Model

Having introduced the three levels of context sensitivity that comprise metacognitive governance, let’s step back and compare the double accumulator model of metacognition to the actor-critic model of reinforcement learning. It should be clear from the description above that the actor is potentially on par with the sensory accumulator, while the critic is potentially on par with the control accumulator. In what follows, these subsystems will be demonstrated to possess a close conceptual proximity, a strong family resemblance.

Recall how the critic subsystem evaluates the performance of actions according to certain control objectives, in particular the tactical and strategic objectives. Epistemic norms, such as
exhaustiveness, accuracy, and fluency, set by the control accumulator, are thus potentially on par with the control objectives set by the critic, as the internalization of these norms is indeed what the control comparator is controlling for. In this respect, one can reinterpret “maximizing the primary reinforcement value” (see above) within the context of metacognition to be a matter of the function of the control comparator to provide the sensory comparator with feedback that will result in epistemically sensitive action.

Consider the act of calculating a sum $\psi$. For $\psi$ to be successful and epistemically sensitive, it must correlate with parameters set by the control accumulator relative to the action, parameters which prescribe an epistemic norm. In this case, let’s stipulate that the parameters are specified to facilitate the performance of an accurate action: $(Acc_i^\psi, ..., Acc_j^\psi)$. Meanwhile, the dynamics of the sensory accumulator performing the act of calculating the sum $(Sum_i^\psi, ..., Sum_j^\psi)$ will be observed and compared by the control accumulator to the dynamics described by its parameters.

In the event of a discrepancy, the control accumulator will produce a feedback signal $\pi$ that is a non-conceptual, analogue signal about the probability of error/accuracy regarding $\psi$, but, for the sake of simplicity, this signal will be represented as a scalar, yielding: $\pi = ((Sum_i^\psi, ..., Sum_j^\psi) - (Acc_i^\psi, ..., Acc_j^\psi))$. This feedback signal will then be sent to the sensory accumulator so that its performance going forward $\psi'$ and beyond ($\psi'' ...$) will better correspond to the control objectives.

$$\psi \rightarrow \psi' = \psi + \pi = \psi + ((Sum_i^\psi, ..., Sum_j^\psi) - (Acc_i^\psi, ..., Acc_j^\psi)) \quad (2)$$

Now, compare the feedback signal produced by the control accumulator to the feedback signal produced by the critic subsystem by contrasting (2) with (1) from above. Of course, two crucial differences emerge, namely how $(Sum_i^\psi, ..., Sum_j^\psi)$ in (2) differs from $S$ in (1) and how $(Acc_i^\psi, ..., Acc_j^\psi)$ in (2) differs from $(C_i^\phi, ..., C_j^\phi)$ in (1). But are they really all so different?

Recall what $S$ represents: the initial feedback value the actor subsystem assigns to the action $\psi$ during performance by comparing the observed dynamics of the system to the expected dynamics: $(f' - f)$. Likewise, the values of $(Sum_i^\psi, ..., Sum_j^\psi)$ are assigned by the sensory accumulator, which are obtained during the performance of the action by making predictions and comparing these predictions to observed dynamics. Meanwhile, as mentioned above,
\((\text{Acc}_i^\psi, ..., \text{Acc}_j^\psi)\) and \((\text{C}_i^\varphi, ..., \text{C}_j^\varphi)\) are functionally equivalent due to how they represent corresponding control parameters.

Thus, the double accumulator model is so conceptually congruent with actor-critic model that it is difficult to distinguish at the cognitive level of description its candidate metacognitive signal \(\pi\), which is reflective of cognitive affordances and noetic feelings, from the non-metacognitive reinforcement signal \(\delta\) described by the actor-critic model. As a result of this strong family resemblance, one can charge Proust’s theory of metacognition with the problem of triviality: if the mechanism behind metacognition is as cognitively simple so as to be equivalent to those behind reinforcement learning, it is deeply unclear how this mechanism implies an advancement in cognition beyond that of basic feats of cognition such as adaptive behavior.

Of course, one might point to either the cognitive process that produces \textit{Sum} (i.e. the mental act) or \textit{Acc} (i.e. the epistemic norm), but either appeal will struggle to solve the problem. With respect to an appeal to \textit{Acc}, here one would need to make the case for how the processes that form these parameters are more sophisticated than the processes that form the parameters of \textit{C}. But considering how they are both essentially control objectives, formed through natural evolution and the individual’s own learning history, this appeal is unlikely to succeed.

Finally, one might appeal to \textit{Sum}, which would boil down to arguing that \(\psi\) (a mental act) is more sophisticated than \(\varphi\) (a bodily act).\(^{22}\) Though the performance of mental acts might indeed require more cognitive sophistication than the performance of bodily ones, this appeal is unlikely to succeed in establishing metacognition as more sophisticated than basic feats of cognition. For one, mental acts are instances of basic cognition. For example, no one would propose that the mental act of memorial recall is particularly sophisticated. This route is thus quite unpromising as it would entail having to explain how such animals, those which are typically denied metacognitive capacities, are able to perform the tasks they can, such as recalling information from memory, without having the cognitive control structures that make such tasks reliable. Therefore, though the double accumulator model is an essential component to the explanation of how noetic feelings are produced, it fails to make clear how this mechanism entails a level of sophistication over reinforcement learning due to its strong family resemblance to the actor-critic model.

The route for shoring up metacognition that ought to be pursued is different from either of the two approaches discussed above, and, in many ways, far simpler. Instead of appealing to

\(^{22}\) This is because of how \textit{Sum} is nothing more than the feedback signal produced while performing \(\psi\), while its counterpart in the actor-critic model \(\delta\) is the nothing more than the feedback signal produced while performing \(\varphi\). Thus, what distinguishes these signals according to their models is just the act that produces them.
conceptual representations to secure a metarepresentational reading of noetic feelings or appealing to the origin of a normativity of an epistemic kind, the route pursued by the present work (both the one advocated here, as well as the alternative that will be discussed afterward) appeals to noetic feelings as conscious forms of control over cognition and behavior (see Chapters 4 and 5).

Thus, the added layer of cognitive sophistication required to make this cognitive control process reflect an advancement over basic feats of cognition (and thus solve the problem of triviality) is the capacity to produce conscious feedback signals. As will be discussed below, evidence for construing mental states characterized by noetic feelings as conscious episodes of executive control over cognition and behavior will suffice in solving the problem of triviality (Chapters 4 and 5). Meanwhile, appealing to noetic feelings as strongly embodied allows us to account for recent empirical evidence of the intricate brain-body mechanism that produces noetic feelings, as well as provides us with essential tools for building the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge (Chapter 6). But before we delve into this, let us return to Proust’s account of noetic feelings and determine the extent to which her theory yields essential ingredients to shore up evaluative metacognition in this manner.

9. How Noetic Feelings Are Distinct From RPEs

Before concluding, I shall examine Proust’s account of noetic feelings and distinguish them from RPEs. I shall demonstrate that the trait that best distinguishes them is their capacity to be accessible for conscious report. To begin, consider how Proust (2013) describes noetic feelings as “functional elements in a metacognitive TOTE loop” (ibid: 22) and functioning as “a pre-specified state of a comparator” (ibidem). Overall, she makes it clear that noetic feelings are nothing less than feedback signals produced by metacognitive monitoring and control (ibid: 19). It is unclear, however, whether noetic feelings are produced by the sensory comparator (see 56-57), the control comparator (see 104), or both. Given that there are references to both, I interpret Proust as suggesting that both comparators are capable of producing noetic feelings. So let’s examine both aspects of this mechanism.

Recall that there are two levels of comparison and two levels of feedback. The first level of comparison is performed by the sensory comparator and this level of feedback informs the system of how well the action and its underlying process are being performed, as Proust says, “under uncertainty”. If the sensory comparator produces a noetic feeling, this feeling will thus likely be the
Chapter 3: Problems Facing the Defense of Evaluative Meta cognition

115

noetic feeling of fluency that informs the cognitive system (as well as the animal, potentially) the degree to which the action is processed quickly and/or adequately according to the system’s own internal standards of performance. In other words, this noetic feeling is a signal that tells the system how well the act is coming along (see 130, 136).

Meanwhile, an additional level of feedback is provided by the control comparator that compares the performance of the sensory comparator and this initial feedback to its own objectives, which, Proust argues, enables the “assessing of one’s uncertainty”, a “different adaptation” from the one needed to “make a decision under uncertainty”. While the production of a noetic feeling of fluency demands only the computing of the so-called ‘balance of evidence’ between competing options, the level of feedback issued by the control comparator is determined by comparing the first level of feedback to control objectives that prescribe epistemic norms, thus producing a signal that, Proust argues, provides, rather than merely carries, epistemic information (see 102-104).

Though Proust does not elaborate on what it exactly means for the feedback signal to provide (rather than merely carry) epistemic information, one can interpret her as saying that the signal produced by the control comparator makes confidence (or its inverse, uncertainty) accessible to the subject. This seems to be what Proust has in mind, since she justifies the maneuver from “merely carries” to “provides” by appealing to evidence from uncertainty monitoring paradigm in comparative psychology (see Introduction), which suggests some animals are able to utilize an uncertain response strategically in order to avoid hard-to-discriminate trials.

If this interpretation is correct, Proust claims that the advanced decision-making behavior observed in some animals in the uncertainty monitoring paradigm (i.e. the ability to opt in and out strategically depending on trial difficulty) is made possible by a feedback signal that makes epistemic information accessible to the animal (see 102-104). Ultimately, this epistemic information is a description of degrees of confidence (or lack thereof) with respect to the animal’s own actions. In the case of the uncertainty monitoring paradigm, this act is often that of perceptual discrimination.

Thus, noetic feelings produced by the control comparator are distinct from noetic feelings produced by the sensory comparator, in that the former are mediators of confidence. As such, noetic feelings of confidence do not merely inform the cognitive system about how the action is fairing, but, moreover, provide the subject herself with “non-conceptual analogue information about the probability of error/accuracy” of her actions in the form of feelings (104).

Thus, two features distinguish Proust’s noetic feelings from RPEs. On the one hand, noetic feelings are argued to pertain to information of an epistemic kind, while RPEs are not theorized
even to carry epistemic information. Second, some noetic feelings are argued to provide, rather than carry, epistemic information, interpreted here as implying personal level (or animal level) accessibility, while RPEs are theorized to operate sub personally. That said, it is not immediately obvious whether these features succeed in making noetic feelings distinct from RPEs.

Concerning the appeal to information of an epistemic kind, this move is argued to be justified by appealing to how actions become evaluated in terms of how fluently they are processed, how quickly or accurately the actions are proceeding according to internal standards, a move which is based on Proust’s proposal that fluency is an epistemic norm. But it is unclear how this process is fundamentally different from the process that the critic subsystem engages in when evaluating the actor’s performance according to its control objectives, since this is also based on evaluating whether actions accurately or quickly conform to internal standards.

As it regards the appeal to the capacity of noetic feelings to provide (rather than merely carry) epistemic information, if the interpretation above is correct and noetic feelings make epistemic information accessible at the personal level, this appeal will succeed in distinguishing noetic feeling from RPEs. This is because the alternative would be to make the case that RPEs are available for conscious report, which no one will be content with. Consequently, the proposal on the table is that noetic feelings ought to be distinguished from RPEs by appealing to them as conscious. This appeal clearly marks a significant break from RPEs and, in virtue of this, succeeds in distinguishing the double accumulator model from the actor-critic model. As a result, light is shed on the way forward. In order to shore up metacognition against the problem of triviality, one will need to appeal to noetic feelings as consciously accessible (see Chapters 4 and 5).

10. Conclusion

As discussed above, the metarepresentation approach, even if successful, will not aid us in shoring up metacognition against the problem of triviality due to how it would need to appeal to conceptual capacities to succeed in securing a metarepresentational construal (either variant), capacities which reflect part of what metacognition aims to explain. Thus, it follows, (if these two are the only approaches), the cognitive role approach is to be preferred.

Proust’s cognitive role approach consists in arguing that noetic feelings ought to be construed as metacognitive due to the crucial role they play in enabling highly flexible cognition that deals in epistemic norms, potentially explaining the emergence of epistemic normativity and,
with it, distinctly human forms of cognition (see Chapter 1). However, in taking the cognitive role approach, the problem of triviality, and its related problems, the problem of metacognitive biodiversity and the problem of honeybee metacognition, will need to be resolved, and hopefully, in such a manner that inroads can be made into solving the ultimate problem of the origin of self-knowledge.

As of now, the cognitive control mechanisms underpinning metacognition, as demonstrated by its strong family resemblance to the control mechanisms for reinforcement learning, are so cognitively basic that it is unclear how an animal mind with the basic tools for reinforcement learning would not also thereby possess capacities for metacognition. If this is the case, the problem of triviality rears its ugly head, also robbing metacognition of its status as a precursor to self-knowledge. As discussed above, the solution to these problems will likely involve an appeal to noetic feelings as characterizing conscious states of executive control over cognition and behavior, hence the approach that will be pursued in the next chapter.
Chapter 4

The Case for Embodied Metacognition:
Shoring up Metacognition against the Problem of Triviality

Abstract

As previously discussed (Introduction), the theory of metacognition faces the problem of triviality: it is unclear how metacognition, as it is often construed, describes anything other than a family of mechanisms and capacities that enable little more than basic feats of cognition. This was made salient above by demonstrating the strong family resemblance of Proust’s double accumulator model, which describes mechanisms of metacognition, to that of the actor-critic model, which describes mechanisms for reinforcement learning (Chapter 3).

In this chapter, it shall be demonstrated how metacognition can be shored up against the problem of triviality by appealing to noetic feelings as strongly embodied. Produced by an the intricate brain-body interface argued for above (Chapter 2), strongly embodied noetic feelings will be demonstrated to be conscious expressions of emotion that enable and facilitate executive forms of control over cognition and behavior. This augmentation to metacognition, dubbed ‘embodied metacognition’, is argued to solve the problem of triviality due to how mental states characterized by strongly embodied noetic feelings represent (in the de re mode of reference) proto states of self-belief. What is potentially more, this augmentation might also provide the necessary tools for constructing a bridging leading from the phylogenetic origin story toward the ontogenetic origin story about self-knowledge that will be argued for in the final chapter (Chapter 6).

1. Introduction

In Chapter 2, I made the case for embodied metamnemonic cognition. This case can be readily extended to include metacognition as a whole in much the same manner that Proust’s metamemory-based theory of metacognition can be extended: by expanding the reach of the
control mechanisms to underpin not only mnemonic capacities, but also capacities for perception, decision-making, and behavior. As such, the cognitive control mechanisms underpinning Proust’s noetic feelings are not too terribly different from the cognitive control mechanisms that underly strongly embodied noetic feelings.

The key difference lies in the degree to which the extracerebral body is implicated in the production of noetic feelings. Within Proust’s account of noetic feelings, the extracerebral body does not play an essential role in explaining how these control mechanisms function, that is, how they achieve the successful control of cognition and behavior (Chapter 1). Thus according to Proust’s theory of noetic feelings, the body plays a much less significant role in this cognitive control process, such that her account might be even categorized as perhaps one of the traditional views discussed above (Chapter 2).

Within the framework of embodied cognition, bodily processes are not insignificant or peripheral components, not addons, but essential parts of the cognitive process, so that any explanation of the target cognitive process will remain insufficient if extracerebral body is neglected (see Shapiro 2019). Thus, in accounting for strongly embodied noetic feelings, I shall argue that the body has an essential part to play in the story of how metacognitive mechanisms enable control over cognition and behavior, and this augmented version, dubbed ‘embodied metacognition’, will be argued to shore up metacognition against the problem of triviality.

Below I begin by reviewing the case for an intricate brain-body interface involved in the production of noetic feelings (see Chapter 2), in particular recent empirical studies that investigate a common mechanism between metacognition and interoception will be discussed again, from which emerges more support for a Neo-Jamesian theory about noetic feelings (see Section 2). In Section 3, the Neurovisceral Integration Model (NVM) will be appealed to in order to interpret the results of these studies, suggesting noetic feelings characterize mental states of executive control over cognition and behavior. It shall be argued that this relationship between mental states and executive function occurs in virtue of how these states index neurocomputational changes (see Section 4). For this and other reasons, these mental states ought to be construed as proto states of self-belief (see Section 5). Finally, some potential criticisms are raised and responded to (see Section 6).
2. Revisiting the Neo-Jamesian Theory of Noetic Feelings

Recall the intricate brain-body mechanism involved in the production of noetic feelings argued for above (see Chapter 2). At its core, embodied metacognition is a family of cognitive control mechanisms, not unlike those described by Proust’s theory of evaluative metacognition. But the cognitive control mechanisms underpinning the production of strongly embodied noetic feelings crucially include processes outside the skull and within the extracerebral body. Thus, questions are raised about the degree to which the extracerebral body is involved in the production of noetic feelings. Below I shall draw upon the Neurovisceral Integration Model (NVM) in order to articulate the exact role of the body in embodied metacognition (e.g., Smith et al. 2017). But before doing this, it will be helpful to review the emerging empirical findings concerning noetic feelings already introduced above, and in so doing, setup the case for embodied metacognition.

Recall from above that pivotal sources of evidence for the crucial role of the body in the metacognitive control process emerges from several recent studies on subjects’ sensitivity to sensory information arriving to the central nervous system from the cardio-respiratory solely system. Often measured by the Shandry heartbeat detection test (1981), though novel tasks that aim at acquiring more precise results are in the process of development (Harrison et al. 2020), interoceptive sensitivity is a measure of how accurately subjects can detect afferent signals sent to the brain by the extracerebral body.

One of the various foci of these studies is on the excitation of a certain class of sensory receptors known as ‘arterial baroreceptors’, which are stimulated by the distortion of the arterial wall during changes in blood pressure. These receptors discharge during the phase of the cardiac cycle known as ‘systole’, in which cardiovascular afferents are sent via the cranial nerves to the central nervous system, in particular to the solitary nucleus. The internal sensation of one’s own heartbeat is thus mediated by the excitation of these sensory receptors (see Garfinkel et al. 2014).

Recall the study conducted by Fiacconi et al. (2016). This study provides us with evidence about what exactly one feels whenever undergoing a noetic feeling.\(^{23}\) Essentially, whatever is felt is

\(^{23}\) To be clear, this study demonstrated a link between cardiovascular afferents and noetic feelings as it regarded only the noetic feelings of familiarity/knowing. Mentioned earlier, I am only concerned with noetic feelings associated with semantic memory, and not noetic feelings associated with episodic memory, such as the feeling of mental time travel (see Michaelian 2016), about which questions regarding their embodiment are not controversial (see Chapter 2). With respect to other semantic noetic feelings, in particular the noetic feelings of confidence/uncertainty, it should not worry us whether these noetic feelings are produced by distinct mechanisms. Recall how Proust’s theory of noetic feelings theorized that noetic feelings of confidence/uncertainty to be the source from which the noetic feelings of familiarity/knowing plausibly spring (see Chapter 1). Moreover, the relationship found between noetic feelings of
crucially linked to the onset of changes in the cardio-respiratory system. After a training stage, in which subjects learn to associate presented images of faces with names, subjects are asked to report whether they experience feelings of familiarity when presented with these images again. But only some of the images were actually presented during the training phase (viz. targets), while some of them were not presented earlier (viz. lures).

If the images were presented at the systolic phase of the cardiac cycle, in which arterial baroceptors are stimulated and cardiovascular information arriving to the central nervous system is at its strongest, the presentation of the face correlated with the reported onset of a noetic feeling of familiarity, a correlation that held for targets, as well as for lures, and regardless of the facial expression (whether emotional or neutral in countenance). Interestingly, this correlation was not found if images were presented at diastole, the other phase of the cardiac cycle, in which cardiovascular information arriving to the central nervous system is at its weakest.

This suggests that a noetic feeling of familiarity can be artificially produced by syncing an image’s presentation to be on, rather than off, heartbeat. Thus, one can infer that, at the very least, interoceptive information arriving from the cardio-respiratory system correlates with the onset of noetic feelings. Plausibly then, this correlation is to be explained by how interoceptive information arriving from the cardio-respiratory system has the potential to produce noetic feelings. In the spirit of the James-Lange hypothesis, it is plausible then that afferent cardio-respiratory signals have this potential to produce noetic feelings in the subject in virtue of carrying information about those extracerebral changes that constitute what is felt whenever a subject feels a strongly embodied noetic feeling.

Let us consider the results of another study that supports this hypothesis. Recall the study conducted by Fiacconi et al (2017), which showed that subjects with improved interoceptive sensitivity will rate their noetic feelings as having higher levels of intensity. Thus, the more acutely one perceives cardio-respiratory afferents, the more acutely one will feel a noetic feeling of familiarity, so that sensitivity to interoceptive information arriving from the cardio-respiratory system predicts the reported intensity of noetic feelings.

One can infer from this that the more sensitive a subject is to cardio-respiratory activity, the more sensitive the subject will be to the production of noetic feelings, that is, the more aroused the subject will be by the noetic feeling. Thus, this study shows a connection between interoceptive sensitivity and the reported intensity of noetic feelings, which provides more empirical support for
the James-Lange hypothesis about strongly embodied noetic feelings. Quite plausibly, what you feel whenever you feel a noetic feeling (i.e. its character) is nothing less than changes produced by cardio-respiratory activity.

This study also demonstrated how subjects reporting higher levels of intensity to their noetic feelings of familiarity will be more likely to describe the stimulus as familiar. One can readily infer from this that the more sensitive a subject is to cardio-respiratory activity, the more influence noetic feelings will exert on the subject’s decision making. In other words, the more intense the feeling of familiarity is (i.e. the higher its degree of arousal), the more familiar a stimulus will appear, and so the more likely the stimulus will be appraised by the subject as familiar.

Combining these conclusions with the discussion above (Chapters 1 and 2), this implies that the more sensitive a subject is to cardio-respiratory activity, the stronger will be the imperative associated with the cognitive affordance prompted by the strongly embodied noetic feeling. For example, say, the cognitive affordance can be described as “Remember!”. It is plausible then that, in the event the subject is extra sensitive to cardio-respiratory activity, the imperative behind the cognitive affordance will be amplified, instituting a stronger demand on the part of the subject to comply with the affordance, stronger than it would be for other subjects who are less sensitive to cardio-respiratory activity.

Thus, if one wishes to provide a sufficient account of how the mechanisms of evaluative metacognition enable and facilitate the control of cognition and behavior, the bodily processes that produce noetic feelings will need to be appealed to, since these processes explain the conditions under which the corresponding cognitive affordances become motivated. That is, at least in the case of creatures like us, and at least in some specific cases, the feedback signals produced by the family of mechanisms underpinning evaluative metacognition are mediated to the personal level by incorporating extracerebral processes, in particular those processes crucially tied to cardio-respiratory activity, processes responsible for translating these cognitive feedback signals into strongly embodied noetic feelings. That is the essence of embodied metacognition.

On the basis of this, the Neo-Jamesian theory of noetic feelings comes into view. The theory holds that what you feel, whenever feeling a noetic feeling, is nothing less than extracerebral bodily activity, felt from the inside, deeply intertwined with cardio-respiratory activity, and hence the subtitle of the present work: “We Feel Our Hearts to Know Our Minds”. Whenever one feels confident or uncertain, feels some piece of information as known though momentarily unable to be recalled, feels that something was forgotten and left behind (the keys, maybe?), feels a friendly face as deeply familiar (what’s his name?), or feels that some claim makes sense or simply cannot
possibly be right, that is, whenever one feels a noetic feeling, one feels bodily activity crucially linked to cardio-respiratory activity that stands in a crucial relationship to neurocomputational activity. Thus, to provide a full account of embodied metacognition, the relationship between extracerebral and neurocomputational activity will need to be outlined, and this is the goal of what remains of this chapter.

3. Strongly Embodied Noetic Feelings Enable and Facilitate Executive Control over Cognition and Behavior

Not only is there evidence for extracerebral activity playing an explanatory role in how cognitive affordances express imperatives at the personal level of description, that is, in the form of strongly embodied noetic feelings, but there is also evidence that strongly embodied noetic feelings enable and facilitate cognitive flexibility and executive functionality. Thus, it is argued below, by appealing to the role that strongly embodied noetic feelings play in enable and facilitating conscious control over cognition and behavior, embodied metacognition, and thus evaluative metacognition, can be shored up against the problem of triviality.

To arrive at this conclusion, key aspects of the empirical evidence from interoception-based research on emotion will need to be introduced (sometimes re-introduced) and thoroughly discussed, in particular the relationship between changes in heart rate variability (HRV) and noetic feelings. Doing so will allow us to draw upon NVM, as HRV is a central component of this model, described as an index of the role that extracerebral body plays an enabling and facilitating executive function. Moreover, drawing upon NVM will allow us to develop an outline of the common mechanism shared between metacognition and interoception argued for in the previous chapter (see Chapter 2, Section 3.4).

First, it is important to remark that your heart is not a metronome. Even if your heart beats exactly 60 times per minute, it is highly unlikely (and would even be unhealthy) for the beats to be exactly one second apart. HRV is a measure of the variation in time between individual heartbeats, so that greater HRV means more variation and lower HRV means less variation in the time in between heartbeats (in ms). Profoundly important for embodied metacognition is evidence from decades of research in cognitive neuroscience that suggests a profound link between more variation in heartbeat frequency and more degrees of cognitive flexibility (for a review, Forte et al. 2019).
To be clear, HRV is a measure of changes in cardiovascular activity, and it is not directly related to interoception. So, it is not as though greater HRV, in its own right, implies a higher probability that subjects will sense their own heartbeat. But due to how interoceptive sensitivity is often measured by determining the degree to which subject can accurately report cardio-respiratory activity, interoception-based studies in emotion often measure HRV in order to draw conclusions about the relative health of subjects’ cognitive and emotional functions. This is because of the established links between these functions and features of HRV.

Explaining the link between HRV and cognitive flexibility is one of the main aims of NVM, which was briefly introduced above (see Chapter 2, Conclusion). Crucially, it describes “an anatomical hierarchy with specific physiological processes that implement bidirectional adaptive control across nested loops of brain circuits”, recently extended to include “a multi-level neural network architecture involving several neurovisceral integration loops” (Smith et al. 2017: 291, 276; emphasis added). This model, originally proposed by Thayer and Lane (2000) describes how, “cardiac vagal tone, indexed by [HRV], can indicate the functional integrity of the neural networks implicated in emotion–cognition interactions” (Park and Thayer 2014: 1).

NVM is thus a hierarchical model of cognitive control based on the Bayesian Brain Hypothesis (Knill and Pouget 2004) and predictive coding architectures (Clark 2013; Feldman and Friston 2010; Friston et al., 2010; Hohwy 2014), describing a cognitive hierarchy made up of eight levels, with the heart positioned just below the bottom-most level, the intrinsic cardiac ganglia. Collectively called the “heart brain” (see Fedele and Brand 2020), the intrinsic cardiac ganglia consist of afferent neurons for receiving information from the other anatomical systems, interneurons for coordinating intra-cardiac reactions, and efferent neurons for transmitting information to the sympathetic, parasympathetic, and central nervous systems. Meanwhile, as one might suspect, the top-most level of the hierarchy is the executive control network, which subserves conscious control over cognition and behavior.

In descending order, the eight levels are: (8) the executive control network, which determines what becomes consciously accessible and held in working memory, (7) the default mode network, which is involved in the categorization of perceptual stimuli; (6) the cortical somatic/interoceptive systems, which regulates current somatic states; (5) amygdala and basal forebrain, which initiates automatic changes to attention and to one’s own somatic states; (4) the periaqueaductal gray and hypothalamic nuclei, which control coordinated skeletal-motor, visceral-motor, and endocrine reactions; (3) brainstem nuclei, which control coordinated cross-organ reactions; (2) spinal and peripheral systems, which coordinate cardiovascular and respiratory control; and finally (1) the intrinsic cardiac ganglia, which coordinates intra-cardiac control.
Within this framework, bidirectional control loops are described as spanning several or more layers, with some control loops spanning even the entire hierarchy, the intrinsic cardiac ganglia included. Information is thus described as flowing from the brain to the heart and from the heart to the brain. The direction flowing from brain to heart is perhaps the most obvious, since executive actions clearly demand metabolic changes mediated by cardio-respiratory activity, but the other direction is just as obvious once introduced.

Afferent signals sent from the intrinsic cardiac ganglia via the vagus nerve to the solitary nucleus in the brain are theorized to constrain multimodal information in the executive control network. This is because of how cardio-respiratory afferent signals convey deviations from physiological states, which will no doubt exert an influence on the subject’s current goals and intentions, as well as exert an influence on which features of the distal environment become salient and engage the subject’s attention. For example, Allan et al. (2016) found that subliminal presentation of emotionally adverse stimuli modulate the influence that sensory noise has on subject’s subjective reports of confidence in perceptual discrimination tasks.

Coupling NVM with the Neo-Jamesian theory of strongly embodied noetic feelings discussed in the previous section, as well as in the previous chapter (Chapter 2), a concrete proposal emerges. Plausibly, the cardio-respiratory afferents that constitute strongly embodied noetic feelings (constitute because these afferents explain the character, arousal, and valence of the feelings) have the potential to shape and constrain multimodal information even at the highest level of the cognitive hierarchy, the executive control network.

Let us now discuss how to make sense of the fact that cardio-respiratory afferents are able to exert a direct influence on executive control, and the implications this has for understanding the relationship between strongly embodied noetic feelings and how they enable conscious control over cognition and behavior. To start, consider how NVM describes the switch from executive control to stimulus control (or executive action to habitual action).

According to NVM, when the dynamic activity at the highest level of the cognitive control hierarchy (viz. the executive control network) fails to match predictions described by its dedicated control mechanisms, conditioned responses enabled by lower levels of hierarchy will instead dominate action selection. In turn, the subject’s behavior becomes less sensitive to its current goals and more sensitive to previously learned associations, producing conditioned responses.

NVM describes an event such as the switch from executive to habitual action as leading to a reduction in HRV. The switch begins with prediction errors at the top of the hierarchy ultimately spurring on activity in the sympathetic nervous system, which has the function of preparing the
body for habitual responses, leading to greater metabolic demand, and so a decrease in metabolic supply (think, fight-or-flight mode), and thus a withdrawal of vagal tone, indexed by a reduction in HRV.

Concerning the switch in the other direction, from habitual to executive action, NVM describes this process as follows. In order for the executive control network to come to dominate action selection, the parasympathetic nervous system will need to first inhibit the sympathetic nervous system. This will result in less metabolic demand, and so an increase in metabolic supply, and thus an increase in HRV. Thus, HRV is a marker of the degree to which the executive control network dominates action selection or struggle to do so, such that greater HRV implies executive control, while lower HRV implies habitual control.

Thus, a draft has been sketched concerning the overarching physiological story of how strongly embodied noetic feelings are part of the process enabling and facilitating executive control. So now let’s apply this model to interpret the results of the studies discussed above (see Chapter 2). Start by recalling Fiacconi et al. (2017). Though Fiacconi et al. did not explicitly measure task-related changes in HRV, it can be calculated nonetheless since they did measure subjects’ interbeat interval (IBI). This is because of how HRV is nothing other than the root-mean-squares of successive differences (RMSSD) in IBIs over a specific time domain, such as what was observed in this study during tasks of metamemory.

Recall that Fiacconi et al. (2017) found that the onset of noetic feelings of familiarity correlated with a decrease in the interbeat interval (IBI) and thus an increase in heartrate. Heartrate acceleration normally causes a decrease in HRV. For example, during exercise and moments of stress, heartrate will increase and HRV will typically decrease. NVM explains this as essentially due to how the layers of the cognitive hierarchy that subserve habitual action and conditioned responses are controlling behavior (see above).

However, contrasting the RMSSD in IBIs for old cues to that of new cues reveals that mental episodes characterized by noetic feelings are associated with greater overall change in IBI during metamemory tasks. This means, noetic feelings are associated with an increase in task-related HRV despite also being associated with an increase in heartrate acceleration that tends to decrease HRV.

Albeit not discussed by Fiacconi et al., this deeply interesting finding can be easily seen by simply contrasting old and new cues in the graph below, specifically contrasting the slopes of IBIs in each condition, starting with stimulus onset and ending with the sixth interval (see Figure 1). Between stimulus onset and the sixth interval, the RMSSD in IBI for old cues was approx. 3ms.
Meanwhile, for new cues it was approx. 2.5ms. Thus, the RMSSD in IBI for old cues is overall greater than it was for new cues, meaning old cues caused greater task-related HRV by an average of 0.5ms between heartbeats.

Though a mere increase of 0.5ms in HRV distinguishes perceiving old cues from new cues, this difference turns out to be far from trivial. For example, Luque-Casado et al. (2015) found that a mere 3ms difference in HRV was enough to distinguish tasks with sustained attention and high executive demand from tasks with little demand on executive function. Plausibly then, mental episodes that involve noetic feelings, not just noetic feelings of familiarity, tend to cause greater task-related HRV relative to episodes in which no noetic feelings are produced.

Again, this is striking given that an increase in task-related HRV was observed despite also the observation of an increase in heart rate acceleration, and, what is more, the increase in HRV was greater than it was for the presentation of new cues. Here, subjects presumably discriminated visual stimuli as novel, which is an action that, of course, depends upon executive function. This 0.5ms increase in task-related HRV for old cues relative to new ones thus potentially demonstrates that a more executively-demanding act is involved in registering the familiarity of a stimulus than that of registering the novelty of the stimulus, due to how this former act is quite plausibly the executive action of attempting memorial recall and so exercising conscious control over mnemonic cognition.

![Fig. 1](image-url)  
*Fig. 1.* Group-averaged and baseline-corrected inter-beat interval (IBI) changes relative to memory cue onset for both old and new cues (trials with recalled information excluded) -- taken from Fiacconi et al. 2017.
Due to this correlation with greater task-related HRV (compared to cases where no noetic feeling is produced but executive action is implied), it is plausible that noetic feelings are an upshot of cognitive control loops that facilitate the executive control network. Plausibly then, if the Neo-Jamesian theory is on the right track, what subjects feel, whenever feeling strongly embodied noetic feelings, is not only an increase in metabolic supply and the introduction of vagal tone, but also the motivation to engage executive functioning, suggesting further that strongly embodied noetic feelings are crucially tied to conscious executive control, perhaps even enablers of it.

Regarding the possibility of the cardio-respiratory system serving as an enabler of executive control, Smith et al. (2017) refer to “a new science of peripheral neurology”, suggesting that the stimulation of the vagus nerve could serve as a possible therapeutic in disorders of cognitive control. They also suggest certain, non-invasive methods, such as slow yoga breathing, that could assist in the therapeutic process through enhancing vagal tone (ibid: 289-290). Put succinctly, within this literature, the variability in heartrate is described as part of what makes the flexibility in cognition possible: “Much of this literature appeals to the increased capacity for flexible response that high levels of high frequency heartrate variability permit” (Sloan et al. 2017; emphasis added).

This empirically established link between strongly embodied noetic feelings and the physiological mechanisms that enable and facilitate executive control is crucial to the present case of shoring up evaluative metacognition against the problem of triviality. Since NVM maintains that variability in cardio-respiratory activity is the foundation for flexibility in cognition and behavior, so much so, it can be read as making a causal-counterfactual claim: ‘no cognitive flexibility without variability in vagal tone’, it is plausible that the mechanism that produces strongly embodied noetic feelings, whose onset correlates with increases in task-related HRV, plays an enabling and facilitating role in conscious control over cognition and behavior. If this is correct, the cognitive control mechanism that produces strongly embodied noetic feelings is clearly more sophisticated than those which enable basic feats of cognition.

To support this further, let us introduce how Smith et al. (2017) describe the relationship between cognitive flexibility and HRV (for more on the NVM, see Thayer and Lane 2009; Berntson et al. 2021; Petzschner et al. 2021).

In other words, higher HRV may index a greater tendency, either in general (i.e., resting HRV) or during a task (i.e., task-related HRV), to assign high precision.
to prefrontal levels of control that are sensitive to goals and context – and these high prefrontal precision estimates will in turn prevent distracting, goal-/task-inappropriate responses to irrelevant stimuli, and therefore improve performance (ibid: 288).

Essentially, Smith et al. maintain that HRV is a measure of how computational priority is assigned within and across the cognitive hierarchy. Greater task-related HRV indicates priority is currently assigned to levels that subserve executive action and control, while lower task-related HRV indicates that priority is assigned to levels that subserve habitual action and stimulus control.

Thus, the results of Fiacconi et al.’s (2017) study, as interpreted to describe HRV, suggest that strongly embodied noetic feelings characterize mental states that emerge downstream from the assignment of computational priority to levels of the cognitive hierarchy that subserve executive action, having even greater HRV than conscious registration and report. Therefore, this amounts to evidence that mental states characterized by strongly embodied noetic feelings are conscious states involving executive control over cognition and behavior.

Thus, the present case for embodied metacognition augments evaluative metacognition by focusing on the degree to which extracerebral processes are involved in production of cognitive affordances, specifically how these bodily processes ensure the resulting mental state will be poised for conscious control over cognition and behavior. Consider again the noetic feeling of familiarity. According to Proust’s theory of evaluative metacognition, a noetic feeling of familiarity occurs when the feature-based representational system produces a cognitive affordance for recalling information from memory (see Chapter 1). Applying the lessons learned from embodied metacognition, the psychological imperative associated with this cognitive affordance ought to be construed as produced in an extracerebral format that is crucial for explaining how subjects exert control over cognition and steer behavior to follow through on the affordance.

At this point, a solution to the problem of triviality facing metacognition has emerged. Essentially, a proponent of evaluative metacognition can argue that the metacognitive feedback signals, noetic feelings, are conscious enablers and facilitators of executive control, feelings which guide the animal’s engagement with robust epistemic acts due to how these feelings ensure behavior

---

24 I have yet to explain the notion of precision appealed to in this passage, which, to do justice to, would require its own section. But, as a shortcut, one can read precision as describing a kind of computational priority, so that assigning high precision to a specific hierarchical level approximately means prioritizing the computations at that level over the computations at others. Ultimately, levels with higher priority (higher precision) have more influence than those with lower priority on determining action selection. More could be said about the crucial details surrounding the process of precision estimation within predictive coding architectures, but this ought to suffice for present purposes (for more on precision estimation, see Clark 2013).
conforms to the objectives described by its metacognitive control structures. Therefore, embodied metacognition warrants the title *metacognition*, not because its products possess metarepresentational structure or involve second-order states, but due to their cognitive role in enabling conscious control over cognition and behavior (i.e. executive control over cognition: *metacognition*).

However, one crucial aspect to embodied metacognition has yet to be explained. How exactly can strongly embodied noetic feelings fulfill this role of enabling and facilitating executive control? The evidence discussed so far suggests noetic feelings have this role due to how they are produced by specific neurovisceral control loops, which bring about the necessary physiological changes that make executive control possible, but it remains unclear how exactly strongly embodied noetic feelings are enablers and facilitators of conscious control over cognition and behavior.

In the next two sections, I shall argue that strongly embodied noetic feelings are capable of enabling executive control over cognition and behavior due to how they characterize mental states that index neurocomputational changes (*Section 4*), mental states that ought to be construed as proto states of self-belief (*Section 5*). As such, subjects who go through the emotional maturation process of revising strongly embodied noetic feelings are effectively engaging in a basic form of conscious self-control. In order to argue for the first claim, I draw, once again, upon those interoception-based studies on emotion and metacognition outlined above, and, in order to argue for the second claim, I discuss basic and generally agreed upon ideas about the nature of belief.

### 4. Mental States Characterized by Strongly Embodied Noetic Feelings Represent (viz. Index) Neurocomputational Changes

In this section, I shall argue that one crucial factor explaining how strongly embodied noetic feelings are capable of enabling executive control of cognition and behavior is how they index neurocomputational changes. In the previous chapter, these changes were referred to as changes in metamnemonic processing fluency. This switch in terms is only meant to reflect that we are no longer dealing with just mnemonic processes, but also processes related to perception, decision-making and behavior, i.e. not merely metamnemonic signals but metacognitive ones. In the next section, I shall argue that, for this and other reasons, mental states characterized by strongly embodied noetic feelings ought to be construed as proto states of self-belief.
Let us start by discussing details of the empirical evidence for the crucial relationship between noetic feelings and cardio-respiratory activity. Meessen et al (2018) showed how subjects with higher levels of resting-state HRV are more accurate in assessing their mnemonic performance, assessments made on the basis of noetic feelings of fluency and confidence (see the cue-familiarity heuristic discussed in Chapter 2, Section 3.3, as well as Proust’s theory of evaluative metacognition in Chapters 1 and 3). Before determining the consequence of the crucial finding, let’s introduce the study.

The study conducted by Meessen and colleagues on the relationship between HRV and metacognitive awareness consisted of two phases, a study phase and a cued-recall phase. During the study phase, word pairs were presented and subjects predicted whether they would successfully recall the words later on, effectively providing an immediate JOL. Otherwise, a variable number of other pairs was presented, after which point subjects made a delayed JOL about the previously presented pair. Thus, Meessen et al. were able to probe both short-term and long-term memory.

Meanwhile, during the cued-recall phase, one pair (one word) was presented alone and subjects attempted to recall the missing pair (the missing word). Before subjects were tasked with these phases, Meessen et al. measured their resting-state HRV via electrocardiogram in order to investigate whether the ability to monitor and predict one’s mnemonic performance correlated with greater resting-state HRV.

Meessen et al. found a positive relationship between greater resting-state HRV and improved metamnemonic awareness consistent across short-term and long-term memory. The greater the subject’s resting-state HRV, the more accurate was their JOLs, suggesting “interindividual differences regarding metacognitive abilities might relate—to some extent—to physiological trait markers of cognitive inhibitory control capacities” indexed by HRV (ibid: 560).

Now, according to NVM, greater cognitive control, in general, correlates with greater resting-state HRV. Recall why this is theorized to be the case. Greater resting-state HRV indicates a greater tendency on the part of the cognitive system to assign computational priority to higher levels of the cognitive hierarchy that subserve executive control, while lower resting-state HRV indicates a greater tendency to assign priority to lower levels of the cognitive hierarchy which subserve habitual action and stimulus control. Thus, it follows from Meessen et al.’s study that greater cognitive control, measured by resting-state HRV, implies more accurate metacognitive judgements (JOLs). Let us get to the bottom of the how and the why of this result.

According to NVM, greater resting-state HRV is not only an index of greater executive functioning, but also an index of enhanced monitoring and regulation in the seventh level of the
cognitive hierarchy, the default mode network (see Barrett and Satpute 2013). Situated below the executive control network, the default mode network supports the essential functions of evaluative metacognition. It is responsible for the detection of task errors (Silvetti et al. 2014), attention to emotion (Smith et al. 2014), self-referential processing (D’Argembeau et al. 2007), and situational appraisal through interactions with long-term memory processes (Wilson et al. 2014). Crucially for embodied metacognition, the default mode network is also described as initiating “autonomic reactions in response to abstract meaning of particular stimuli… in which the perceptual properties of those stimuli are themselves unassociated with metabolic demand” (Smith et al. 2017: 287).

Consider how crucial this network is for the control mechanisms involved in the production of strongly embodied noetic feelings. Not only is this network responsible for task-related error detection, but it is also responsible for producing autonomic responses to abstract stimuli, such as word pairs, people’s names, and the groupings of pixel densities, i.e. the various stimuli deployed by cognitive psychology for producing noetic feelings. Thus, it is plausible that greater resting-state HRV is a marker of enhanced embodied metacognition, possibly due to an enhanced potential to produce more accurate noetic feelings.

Recall from our discussions above that noetic feelings are crucially tied to cognitive affordances, and cognitive affordances are theorized to be mental representations with correctness conditions structured by specific vehicle properties of neurocomputational changes, such as the firing rates of neuronal populations. Staying within the NVM framework, it is plausible then that the enhancement in the default mode network, which accounts for improved accuracy of JOLs, can be itself explained by appealing to an improved mechanism for producing relevant cognitive affordances as strongly embodied noetic feelings.

Thus, the conclusion that can be drawn from Meessen et al. (2018) is this. It is in virtue of some property of mental states characterized by strongly embodied noetic feelings that a relationship obtains between better cognitive control and more accurate JOLs. Plausibly, this property is something not too terribly divorced from a kind of fidelity in translation, wherein the source language is neurocomputational activity and the target language is bodily activity, specifically (given Fiacconi et al. 2016, and Fiacconi et al. 2017) cardio-respiratory activity: the better the translation, the more accurate the noetic feeling, the more accurate the JOL.

Recall from Proust’s theory of evaluative metacognition (see Chapter 1) that the valence of noetic feelings is a product of better-than or worse-than anticipated performance of cognitive processing, while the arousal of noetic feelings is a product of the degree to which this performance is better-than or worse-than anticipated. What follows from the Neo-Jamesian theory of strongly
embodied noetic feelings is that valence and arousal of noetic feelings are ultimately products of extracerebral activity, in particular, as Meessen et al.’s study and Fiacconi et al.’s (2017) study demonstrates, activity indexed to changes in HRV.

If this is correct, the enhanced metacognitive performance observed by Meessen et al.’s study is to be explained by appealing to how subjects with greater resting-state HRV tend to have feeling-based cognitive affordances that more accurately represent neurocomputational changes, so that subjects with greater resting-state HRV undergo noetic feelings that more accurately represent the success (or failure) of their executive actions, serving as more effective guides to their behavior. Therefore, it is plausible that improved accuracy in evaluative metacognition, as measured by Meessen et al.’s study, is plausibly due to the production of strongly embodied noetic feelings that more accurately represent vehicle properties of neuronal populations.

If this is correct, it means subjects with greater resting-state HRV undergo noetic feelings with both valence and intensity that more accurately represent degrees of better-than/worse-than predicted performance of neurological computations due, in part, to having a more intricately fine-tuned heart-brain relationships. What is potentially more, it also means, for people with more intricately fine-tuned heart-brain relationships, those essential neurocomputational changes are more accurately represented in the embodied components to the feeling.

To put that succinctly, having a more intricately fine-tuned heart-brain relationship helps the heart tell subjects about the details of their minds. Consequently, mental states characterized by strongly embodied noetic feelings are conscious states of executive control over cognition and behavior, in which patterns of bodily arousal tied to cardio-respiratory activity represent patterns of neurocomputational activity.

This point about the representational structure of strongly embodied noetic feeling is crucial to the present case. But recall what exactly ‘represents’ means in this context. Following Proust’s theory, noetic feeling do not represent in the de dicto mode of reference (see Introduction and Chapter 1). Rather, noetic feelings represent vehicle properties, in the sense of indexing them (see above, also Shea’s account of de re representation). Specifically, this form of representation involves such strong informational covariation that ‘representings’ (noetic feelings) can fulfill key aspects of the functional role of ‘representeds’ (neurocomputational changes) even (and especially) in their absence (see Recanati 2012).

One might wish to say, as Carruthers (2017) does, that noetic feelings represent features of the external world and not features internal to the animal’s own cognition, but this would not pose any problem for this account of the representational structure of noetic feelings. Carruthers
construal of the content of noetic feelings is reflective of what he thinks might be potentially available at the animal level of description. He is arguing for the *de dicto* mode of reference (see *Introduction*), which is orthogonal to the case for the *de re* mode.

The central claim of this section is that noetic feelings and their cognitive affordances are conscious, a claim Carruthers has recently endorsed (see Carruthers and Williams 2019). The claim is not, of course, that vehicle properties of neuronal populations are conscious; rather mental states characterized by strongly embodied noetic feelings possess strong informational covariation with neurocomputational states (and so represent them in the *de re* mode of reference).

In the next section, it will be argued that mental states characterized by strongly embodied noetic feelings ought to be construed as proto states of self-beliefs in part due to how they represent vehicle properties of neuronal populations, that is, regardless of whether these states represent these states as features of the external world or as features internal to cognition. Below the argument for construing these mental states as proto states of self-beliefs turns on rational considerations about what would qualify a mental state as approximating a belief state.

5. Mental States Characterized by Strongly Embodied Noetic Feelings Ought to be Construed as Proto Self-Belief States

In this section, I propose construing mental states characterized by strongly embodied noetic feelings of confidence (and its inverse, uncertainty) as well-suited to transform into full-blooded states of self-belief. Once, as the present work assumes (see *Introduction*) the required societal pressures are in place for the development of metarepresentational capacities that would, in turn, make possible the representation of beliefs as such (for more details on this process, see *Chapter 5*), mental states characterized by strongly embodied noetic feelings of confidence can transform into full-blooded mental states characterized by self-belief.

However, in order to argue for this, I do not appeal to empirical evidence related to the production of noetic feelings, as I have above. Instead, I appeal to considerations about which features would likely constitute approximations of full-blooded self-beliefs and present the mental states characterized by strongly embodied noetic feelings as approximating those features.

At first blush, full-blooded self-beliefs states are likely the products of robust internal models that describe aspects of the subject herself as aspects of the subject herself. Of course, this
is *de dicto* mode of reference, which entails capacities for metarepresentation (see Introduction), and such a condition would be too much to demand from a potential precursor. So the first aim is merely to show how mental states characterized by strongly embodied noetic feelings are likely produced by models that approximate descriptions of the subject herself as the subject herself.

In the first instance, it will need to be demonstrated how mental states characterized by strongly embodied noetic feelings describe aspects of the subject herself (without the ‘as’ bit). This makes sense, obviously, given how these mental states are conscious emotional states of executive control, which exhibit robust informational correlations with patterns of neurocomputational activity, which, in turn, describe the adequacy of executive actions in conforming to stored profiles of performance (see Chapter 1).

In other words, though mental states characterized by strongly embodied noetic feelings are not produced by internal models of the subject herself as the subject herself, they are produced by models that approximate these. Thus, mental states characterized by strongly embodied noetic feelings are crucially analogous to self-belief states, in part, due to how they are caused and correlate with internal descriptions of the subject’s own actions and how they inform the subject about aspects of herself (features of her actions) – the critical difference being, of course, that these mental states do not inform the subject about aspects of herself as aspects of herself.

As such, the proposal is that, albeit not full-blooded self-belief states, mental states characterized by strongly embodied noetic feelings ought to be construed as *proto* self-belief states. In this regard, mental states characterized by strongly embodied noetic feeling of confidence and uncertainty are of particular interest. Plausibly, these emotional states reflect nothing less than states of self-confidence and states of self-doubt, respectively. Let examine this proposal further by introducing two generally agreed upon criteria for any belief state, about the self or otherwise.

One crucial criterion distinguishing a mental state as a belief state is the possession of the concept of belief, namely that the corresponding mental state discloses a perspective on the world, one which aims at truth (see Davidson 1982). Another crucial criterion distinguishing belief states is the so-called ‘mind-to-world direction of fit’, so that, in the event the mental state misrepresents, it is the mental state that must fit to the world, and not the other way around (see Searle 1983).

As already discussed in detail above (Chapter 1), conceptual capacities fall outside the realm of non-human animal cognition (and, with them, representations that aim at truth), so it would be a mistake in our case to demand mental states characterized by strongly embodied noetic feelings meet the first criterion as it currently stands. That said, while the representational products of the animal mind do not aim at truth, their mental representations do, nonetheless, aim at accuracy.
Recall that this is due to how the activity in the control comparator is a function of cognitive fluency, and fluency, when not misguided, approximates accuracy (see Chapter 1). So mental states characterized by strongly embodied noetic feelings of confidence can be construed as aiming at accuracy and this would not entail a level of cognitive sophistication that outstrips the minds of non-human animals, while also approximating the sophistication required by the concept of belief.

Now, consider the direction-of-fit exhibited by mental states characterized by strongly embodied noetic feelings of confidence. In the event the cognitive system misrepresents confidence, it is the state (or features of the state) that ought to change in order to resolve the conflict and ensure the subject’s adaptive fitness, not the other way around. In other words, if one feels confident in one’s ability to perform an action, and one fails to perform it well, more often than not, it ought to be the confidence level that changes, and not some facts about how the action should be performed.

Going off of these considerations then, mental states characterized by noetic feelings of confidence have mind-to-world direction of fit similar to that of belief. Therefore, it is plausible that mental states characterized by strongly embodied noetic feelings of confidence have the same direction-of-fit as belief, namely mind-to-world, which thus makes these states belief-like. Though this is by no means sufficient, it is a necessary hurdle that needs to be cleared.

Assuredly, pressures exist on the part of the individual to ensure this process of revising misplaced self-confidence occurs, pressures which are independent of any societal pressures to do so. For any animal that fails to revise its confidence in its abilities, and instead persists with misplaced confidence, will not survive for long. Unless, of course, it continuously finds itself in very hospitable and forgiving environments. But if its world is anything like ours, such luck is bound to run out eventually.

What is more, subjects thus endowed with the brain-body mechanism responsible for producing strongly embodied noetic feelings might be in the position to entertain proto self-belief states. These subjects can be said to deal in proto self-beliefs, internally, in quasi-individual terms, without thereby demanding they possess the concept of belief. Effectively, we are now dealing with intraindividual regulatory practices that approximate updating beliefs in light of evidence and counter evidence. Consider how this could work.

The subject might start off feeling confident in her ability to perform some action. This feeling of confidence is mediated by the monitoring and control mechanisms of embodied metacognition through the production of a strongly embodied noetic feelings of confidence. Next, the subject performs the action but feels a lack of fluency with respect to this performance.
Likewise, the feeling of low fluency is mediated by the mechanisms of embodied metacognition. As a result, she comes to feel differently about her ability to perform the action going forward, comes to feel a noetic feeling of uncertainty, likewise mediated by embodied metacognition. Thus, her confidence levels will tend to become revised in accordance with the subject’s actual abilities after registering feelings of low fluency, thus approximating how beliefs tend to become revised after learning of counterevidence.

One might ask whether it is the subject herself that is revising her confidence or whether it is some subsystem of the subject’s cognition doing the interesting work here. Though, it does not really matter all too terribly much either way, since the present case does not depend on the subject having a robust form of self-awareness (it is assumed that this arrives later, once the societal pressures are in place), but some conclusion about who or what is doing the revising might be arrived at by considering the empirical evidence above which shows how strongly embodied noetic feelings enable and facilitate executive control over cognition and behavior. This suggests it is the subject, and not one of her subsystems, doing the revising.

This process of proto self-belief revision thus sounds, as it should, like an emotional maturation process rather than a reasoning process, but nonetheless reflects a plausible case for how proto self-belief states might go about changing to fit the world rather than the other way around in a manner that aims at truth. So this emotional process is deeply belief-like in the various senses just articulated, pertaining to the self (one’s abilities), approximating truth, and needing to be revised to fit the world (rather than the other way around) in the event of error.

Not only does this revising of proto self-beliefs of confidence occur intra-individually, but it could potentially become enforced by other individuals within the subject’s socio-cultural environment. Subjects who revise self-confidence in their abilities to better match the actual success of these abilities could be rewarded by others in various ways (e.g. by being granted social roles), as well as punished in various ways for failing to do so (e.g. by being denied social roles). This is a crucial point, and it will need to be expanded upon in the final chapter (Chapter 6).

Ultimately then, this explains how strongly embodied noetic feelings of confidence enable and facilitate executive control of cognition and behavior. Plausibly, mental states characterized by strongly embodied noetic feelings of confidence (and uncertainty) reflect proto self-belief states, and so entertaining them and dealing in them internally (e.g. in the emotional maturation process just illustrated) is nothing less than exercising a rudimentary form of self-control over what one believes about oneself and how one behaves in light of those self-beliefs.
6. Response to Possible Criticism

Now that the case for embodied metacognition has been thoroughly discussed, let us discuss two key worries with this proposal. An initial worry is that one might have concerns whether strongly embodied noetic feelings possess sufficiently robust representational structure to explain the emergence of proto self-belief states, while another worry is that an appeal to the extracerebral body and emotional experience in general is adding only a superfluous dimension to a central and cerebral process.

To begin, recall how it was discussed above that the best way of making sense of the representational structure of noetic feelings is the notion of indexing, one which qualifies their mode of reference as slightly more robust than mere statistical co-variance. Like an index in a book, noetic feelings can serve as pointers for neurocomputational changes even if the absence of these changes. Let us consider empirical evidence for this claim.

Introduced above (Chapter 2), Goldinger and Hansen's (2005) study showed how subliminal vibrations could cause subjects to report illusory feelings of familiarity when faced with novel stimuli. They showed this by stimulating embodied components that typically accompany the noetic feeling in the absence of the pattern of neurocomputational activity that would normally start the causal change that ends in the production of the feeling. Meanwhile, Fiacconi et al. (2016) observed similar evidence of the representational structure of noetic feelings when presenting visual stimuli synchronized with the systolic phase of the cardiac cycle. Thus, these studies showed how noetic feelings can misrepresent, which should serve as strong motivation for construing them as having sophisticated representational structure with patterns of neurocomputational activity comprising their correctness conditions.

Concerning the second worry, it is paramount to see just how extracerebral processing is not a mere addon to evaluative metacognition, but rather plays a crucial role in shaping and constraining the parameters of the underlying mechanism. Evidence for this comes once again from Goldinger and Hansen study which showed how old stimuli, when perceived under the influence of subliminal vibrations, were less likely to induce noetic feelings. Plausibly then, the combined effect of having the pattern of neurocomputational activity together with the production of atypical, bodily afferents attenuated the feeling of familiarity, causing subjects to mistakenly report old stimuli as novel.
Recall Allen et al.’s study (2016) that produced similar results. Also recall how these researchers explain this inverted relationship as a matter of how unexpected bodily arousal can counteract the influence of cognitive biases. Such biases, Allen et al. theorize, serve as computational priors, that modulate the influence of sensory processing on conscious experience. Thus, if Allen et al. are correct, it means Goldinger and Hansen’s study showed, not only that noetic feelings have the potential to misrepresent neurocomputational changes, but also how bodily processes integral to the production of noetic feeling can potentially counteract priors about neurocomputational fluency. Under the maligned influence of the subliminal buzzing, higher neurocomputational fluency may come to mean unfamiliar, while lower neurocomputational fluency may come to mean more familiar. Allen et al. see results like these as “motivat[ing] a revised view of metacognition as incorporating [priors] about both physiological states and the precision of actual sensory inputs” (ibid: 7; emphasis added). In summation, the extracerebral body shapes and constrains metacognitive experience and so is, by no means, superfluous to explaining the details of this metacognitive process.

7. Conclusion

Here in the conclusion, I would like to summarize this chapter and introduce the central topic of the remaining two chapters. This chapter primarily concerned how to shore up evaluative metacognition against the problem of triviality. This was done by appealing to an intricate brain-body interface that produces strongly embodied noetic feelings. This has the advantage over the metarepresentation approach of not relying upon such cognitive sophistication that would preclude its role as a precursor to self-knowledge (i.e. conceptual capacities). It also has the advantage over the previous cognitive role approach in that it is clear how this this brain-body interface, as well as the feelings it produces, cannot be so easily accounted for by appealing to relatively simple cognitive mechanisms, like those underpinning reinforcement learning.

First, strongly embodied noetic feeling were argued to be conscious markers of executive control, not only motivators (and demotivators) of executive action, but also crucially part of the process that enables executive functioning. Second, the role was argued to occur in virtue of how the mental states characterized by strongly embodied feelings index neurocomputational changes. Third, due to this representational structure and other considerations about the nature of belief, the mental states characterized by noetic feelings were argued to be proto states of self-belief. In
the final chapter, these proto states of self-belief shall be argued to play a crucial role in plausibly building a bridge between the phylogenetic and ontogenetic origin stories about self-knowledge.

In the next chapter, we shall discuss the problem of triviality in another light, one that is informed by recent trends in cognitive science, namely the Bayesian Brain Hypothesis and the theory of predictive processing. Here, we shall discuss an alternative to the current proposal offered by Meyniel et al. It will be demonstrated how their account offers a similar solution to the problem of triviality, as well as related problems, the problem of metacognitive biodiversity and the problem of honeybee metacognition. However, it will remain nonetheless unclear how Meyniel and colleague’s theory has the tools to account for the emerging empirical evidence about the role of the extracerebral body in the production of metacognitive feedback signals and unclear whether it has the tools to secure a bridge between the phylogenetic and the ontogenetic origin stories about self-knowledge. This point will be revisited in the final chapter (Chapter 6), where we shall discuss not only reason for preferring the present account over Meyniel and colleague’s, but also how embodied metacognition might well provide the required tools for building a bridge between the phylogenetic and ontogenetic origin stories about self-knowledge.
Chapter 5

Evaluative Metacognition and the Predictive Processing Framework: Grappling with the Origin of Self-knowledge

Abstract

The predictive processing framework (Hohwy 2013; Clark 2016) and the Bayesian Brain Hypothesis (Knill and Pouget 2004) offer researchers across the sciences of brain and behavior an exciting and illuminating approach to understanding fundamental aspects of the mind, such that now a growing body of research applies it to the domain of metacognition (e.g. Sherman et al. 2015; Sherman and Seth 2021; Fleming 2021; Hu et al. 2021), in particular by operationalizing confidence as a product of precision estimation (e.g. Meyniel et al. 2015a, 2015b). However, precision estimation is routinely regarded as occurring ubiquitously through the cognitive hierarchy (see Seth 2015), so that confidence, so construed, would then be produced and utilized by practically every layer within the hierarchy (see Friston et al. 2015). But this construal runs the risk of trivializing metacognition, since its explananda are sophisticated cognitive mechanisms, crucially related to reflection and executive function, which could potentially serve as crucial phylogenetic ingredients in the origin of self-knowledge (e.g. Smith et al. 2003; 2012; 2019). One approach that could be used to shore up this metacognition with the predictive processing framework appeals to the global workspace theory to argue for metacognitive confidence signals as conscious (see Meyniel et al. 2015a). I argue this approach succeeds in solving the problem of triviality, as well as related problems, the problem of metacognitive biodiversity and the problem of honeybee metacognition; however, it is unclear whether this appeal can explain recent evidence from interoception-based studies on metacognition or whether it is sufficient to secure metacognition as a crucial phylogenetic ingredient in the origin of self-knowledge.
1. Introduction

Recall from above (Introduction) that metacognition is thinking about thinking, but to say anything more nuanced than this would ultimately force one to take a side in the debate over what metacognition is and how it ought to be construed. Recall further there are two prominent accounts of metacognition. On the one hand, there are so-called ‘mindreading’ accounts, in which the subject represents someone, like herself, as having a mental state and uses this representation to draw inferences about what the person believes or how they will behave. Another way to talk about this account of metacognition is to refer to it as ‘metarepresentational metacognition’ (see Introduction). Crucially, the capacity to represent representations as representations is routinely held to be essential for explaining distinctly-human intelligence, in particular explaining how humans have come to know that they know (meta-cognition) and thus acquire self-knowledge (see Carruthers 2000; Clark 2000; Proust 2013; Bermúdez 2004; Thagard 2005). As such, making sense of how metarepresentational metacognition emerged in nature reflects a proverbial holy grail of cognitive science, as well as the philosophy thereof.

However, mindreading will not be discussed in any more detail here; rather, for reasons that will become clear, this chapter will focus on the other prominent account of metacognition called ‘evaluative metacognition’ (henceforth, ‘metacognition’). As previously discussed, metacognition consists in the deployment of capacities for monitoring and controlling cognitive operations through the production of either conscious or non-conscious so-called ‘noetic feelings’ that motivate and guide behavior for the purpose of mitigating and coping with uncertainty (Dunlopsy and Metcalfe 2008). The literature on evaluative metacognition is vast, stretching back decades (see Hart 1967), and continues to be an exciting and informative theory for understanding various explananda across the sciences of brain and behavior (see Smith and Beran 2021). Most crucially for our discussion, evaluative metacognition has been theorized to serve as a potential precursor to mindreading (see Introduction).

Recently, another equally exciting and illuminating theory enjoying prominence across the brain sciences called ‘predictive processing’ (also called ‘predictive coding’; Hohwy 2013; Clark 2016), which sheds light on all things cognitive (i.e. action, perception, emotion, reasoning, decision-making), has been applied to the domain of metacognition in an effort to expose and make sense of how these metacognitive capacities for monitoring and control emerged (see Fleming 2021). Predictive processing is a general theory about how the brain makes the mind, one which is based on the Bayesian Brain Hypothesis (Knill and Pouget 2004), essentially holding that
Chapter 5: Evaluatorc Metacognition and the Predictive Processing Framework

143

the brain is in the business of generating predictions of its incoming sensory input by deploying generative models that describe the hidden causes of sensory information.

The brain engages processes for both perception and action (perceptual inference and active inference, respectively) that approximate Bayesian inference by issuing predictions based on prior beliefs (subpersonal descriptions of sensory causes), acquiring and weighing sensory evidence, and subsequently updating beliefs in light of evidence to generate new, posterior beliefs. Crucially, these processes unfold hierarchically, so that higher levels of the cognitive hierarchy issue predictions that describe the activity in the lower levels, which then, in turn, send feedback signals about the accuracy of those descriptions in the form of prediction errors, feedback that travels back up the hierarchy, where it is used to revise and update predictions. Much more could be said about the details of how this works, but that account should suffice for present purposes.

That said, our discussion will not concern predictive processing itself so much as it will concern how to integrate metacognition into the fold of predictive processing. At first blush, an immediate worry appears around how best to do this. In particular, it is unclear how the deployment of capacities for monitoring and controlling cognitive operations is any different from the basic business of the brain: each level of the cognitive hierarchy monitors and controls cognitive operations of the lower levels aimed at mitigating uncertainty through the production of feedback signals. So it would look like cognition is just metacognition through and through, that metacognition is just how cognition works. This is deeply problematic though because, if correct, metacognition would be rendered trivial, as it would not describe sophisticated cognitive capacities beyond that of basic feats of cognition (see Chapter 3). Also, it would fail to help us make sense of how to solve the ultimate problem, the problem of the origin of self-knowledge.

The details of this problem will be laid out below, but let us step back and reintroduce the theory that has emerged from decades of research in cognitive and comparative psychology about how metacognition works, which will also serve to introduce the problems that lie ahead. For the moment, however, let us ignore what predictive processing would have to say about this and instead pretend for a moment that this theory does not exist.

Recall that the story of metacognition has its origin in cybernetics, that having a cognitive system is a matter of having the means to monitor and control behavior to ensure the survival of the organism, whose fitness crucially depends upon capacities of the system to adjust behavior whenever some part of the process goes astray (see Chapter 1). In order to be in the best possible position to achieve this, an organism would benefit from having a cognitive system that produces, stores, and updates models that describe the performance of behaviors, a system which is able to
compare the actual performance of any behavior to its own internal standards, in order either to change the performance to conform to them or to change its models to conform to the performance. So far, this story sounds a great deal like the one told by theorists working in predictive processing (see Friston 2019).

Sticking with the standard story of metacognition, in the event of a discrepancy between actual performance and the description in the internal model, a feedback signal communicating the degree of discrepancy is produced and utilized by the cognitive system to update internal models in light of this discrepancy in order to achieve better subsequent performance and ensure the organism’s survival. The mechanism that monitors and controls cognition through the comparison of performance to models and the production of feedback signals is a metacognitive mechanism due to how it concerns, or is about, cognition, and any organism that can adapt behavior to conform to this process possesses metacognitive capacities. So goes the standard story.

Since the distinctly human capacities for mindreading involve, to some crucial degree, capacities for thinking about thinking, cognizing about cognition (either one’s own cognition or that of others), metacognitive capacities for monitoring and controlling cognitive operations (i.e. cognizing about cognition) – which guide subjects in determining whether perception, memory, decision-making, and reasoning are accurate – could plausibly be the source from which these distinctly human capacities spring. This is the guiding thought behind research into metacognition. Thus, evidence of similar capacities, enabling similar behaviors, and underpinned by similar capacities in non-human animals is believed to evidence the role of metacognition as a phylogenetic precursor to capacities for self-knowledge.

Let us remember the theory of predictive processing and ask now, what is wrong with this story? Clearly, the problem is that if this is the correct construal of what metacognition consists in, then quite possibly every living organism possess metacognition, which means metacognition does not reflect a sophisticated advancement over cognition, let alone a precursor to mindreading and self-knowledge (see Introduction). This is because the story above is nothing less than the story predictive processing has to tell us about what cognition consists in (see Hohwy 2013; Clark 2016).

That said, as was mentioned earlier (Introduction), evidence from comparative psychology makes it clear that metacognitive capacities for the monitoring and guiding of behaviors effective at mitigating uncertainty are not equally distributed across the animal kingdom. It is simply not the case that all living organisms possess capacities for metacognition in the same manner that other living organisms do. Explaining this phenomenon, while avoiding the problem of triviality, is effectively solving the problem of metacognitive biodiversity.
For example, pigeons, who are excellent at reward-based associative learning, struggle to produce behaviors for mitigating uncertainty (Sutton and Shettleworth 2008); and dogs, those lovely friends of ours, who are so perceptive and socially creative, struggle to mitigate uncertainty as well (McMahon et al. 2010). Meanwhile, Old-World monkeys (e.g. Hampton 2001), apes (e.g. Call 2010), and dolphins (e.g. Smith et al. 1995) have displayed uncertainty mitigating behaviors that succeed just as well as those produced by adult humans, and they are even susceptible to the same sorts of illusions of metacognitive accuracy as humans (Ferrigno et al. 2017). So metacognition cannot simply be the day-to-day business of cognition. Something has to give.

In what follows, I begin by laying out the framework for animal metacognition with the aim of making it clear why this research is so important for research on predictive processing (Section 2). Thereafter, I discuss the details of the evidence that demonstrates metacognition in animals, how it cannot be reduced to mere associate learning, all the while walking us through the controversy, already introduced (see Introduction), about how best to construe metacognition (Section 3). This is essentially the problem of metacognitive biodiversity.

This discussion will lead us to a case that acutely demonstrates the problem, namely the case for honeybee metacognition, presenting a problem in its own right (Section 4). The possibility of honeybee metacognition is problematic because at first glance it would seem that if honeybees have capacities for metacognition that are as robust as those that humans have, then metacognition cannot clearly reflect a crucial precursor to self-knowledge, simply given the fact that humans and honeybees are separated by over 600 million years of evolution.

Recall that the controversy is about providing an account of metacognition that does justice to evidence coming out of comparative psychology while at the same time warranting the ‘meta’ prefix by providing an account of metacognition that clearly reflects a sophisticated advancement over basic feats of cognition (see Introduction). This is the problem of triviality discussed in detail in the previous chapter (Chapter 4). There the problem was solved by augmenting evaluative metacognition to become embodied metacognition, which produces strongly embodied, conscious noetic feelings that enable and facilitate executive control over cognition and behavior. However, shoring up metacognition is also about providing an account of that makes it is clear how it can serve as a significant phylogenetic precursor to capacities for acquiring self-knowledge (see Lay Summary and Introduction). This is the problem of the origin of self-knowledge, but it will not be thoroughly discussed until the final chapter (Chapter 6).

Thus, the main aim of the present chapter is to discuss an alternative approach to the one offered by the present work, presented in the previous chapter, one based on the theory of
predictive processing (Section 5). The guiding thought within this framework is that distributional confidence (or a precision estimation), approximating Bayesian confidence, is read out by neural circuits to produce scalar representations of summary confidence that are then broadcast globally to become conscious subjective confidence (e.g. Meyniel et al. 2015a, 2015b).

In other words, Meyniel et al. propose a similar solution to the problem of triviality facing metacognition, namely that of appealing to conscious feedback signals. As will be demonstrated below, their account can thus solve the problems of metacognitive biodiversity and honeybee metacognition in a similar manner as the current proposal would. But their account differs from the one proposed here in that it does not appeal to the extracerebral body to account for the production of conscious feedback signals. Rather, Meyniel et al.’s account is conceptually similar to traditional views discussed earlier and so ought to be construed as one (see Chapter 2). That said, it will be introduced in this chapter (and discussed in more detail in the next) how Meyniel et al.’s traditional view struggles to account for the recent empirical evidence about the role of the extracerebral body in the production of metacognitive feedback signals, as well as struggles to account for how metacognition can be construed as an origin for self-knowledge.

In the next chapter (Chapter 6), I provide a framework for how to transform Meyniel et al.’s theory into an account of embodied metacognition. There, I argue that embodied metacognition, rather than a non-embodied account, is better able to explain recent evidence from interoception-based studies on metacognition and better suited to make inroads into the problem of the origin of self-knowledge due to evidence already discussed above and other considerations about what the payoff might be in having strongly embodied noetic feelings.

2. The Problem of Metacognitive Biodiversity: Why it is Important for Predictive Processing

As already stated, I will be discussing the issues that arise when integrating metacognition into the predictive processing framework. Essentially, metacognition needs to be construed within this framework so that its underlying mechanism is still on track to explain sophisticated forms of cognition that imply reflective or executive capacities, as well as potentially serve as a crucial phylogenetic origin of capacities for self-knowledge. The reason why metacognition must be construed in this way is because of the extensive evidence that has emerged in comparative psychology over the past twenty years about the nature of animal metacognition, evidence which
provides a strong case for the possibility of a non-conceptual, non-metarepresentational, procedural know-how and possibly conscious form of metacognition that is nonetheless distinct from the ubiquitous cognition-about-cognition described by the predictive processing framework at every level of the cognitive hierarchy. Let us discuss this evidence now.

Two major paradigms, or families of paradigms, have been developed to test whether animals have capacities for metacognition: the information-seeking paradigm (e.g. Beran et al. 2013) and the uncertainty monitoring paradigm (e.g. Smith et al. 2003). Though it has been pointed out that the information-seeking paradigm has certain advantages over uncertainty monitoring (Marsh 2019), I shall concentrate exclusively on uncertainty monitoring. This is not only because of its immense importance within comparative psychology, but also because of its potential to expose both continuity and conflict with the theory of predictive processing, which is fueled by a cognitive economy whose currency is Bayesian approximations of uncertainty. So let us now turn to both the most celebrated and controversial approach to acquiring evidence for the role of metacognition in explaining the phylogenetic origins of self-knowledge.

2.1 Details Surrounding the Uncertainty Monitoring Paradigm as Evidence for Metacognition in Animals

Though introduced earlier (Introduction), the uncertainty monitoring paradigm have yet to be thoroughly discussed. To begin, notice how humans are obviously aware of their confidence and uncertainty. When faced with difficult decisions, we tend to introspect and consider what we know. Is what I know right? How confident am I that this is correct? Questions like these typify Socratic questioning that was used to introduce the present work (Lay Summary). At its core, the uncertainty monitoring paradigm is designed to probe whether non-human animals engage in a similar introspective process when faced with difficult decisions.

Practically all experiments based on this paradigm have the following setup in common. Both human and non-human subjects are presented with a battery of cases to probe their ability to perceptually discriminate. The chief task is to discriminate stimuli of a specific modality along certain (often polar) dimensions: high vs low, long vs short, dense vs sparse, located above or below a threshold. Subjects make their discrimination known by often pressing one of two buttons, one for either category. For correct responses subjects receive rewards, while incorrect responses result in timeouts in which no reward can be obtained.
Regarding timeouts, it is helpful to compare these to loading screens in videogames, specifically how after the avatar dies, the loading screen appears, and one has to wait an insufferable amount of time before trying again. Here, one often feels cheated. Thinking about timeouts in this way make clear how they can be somewhat punishing and subjects will want to avoid them.

As the experiment goes on, specific trials are introduced that are, by design, extremely difficult and in some cases, even impossible to discriminate correctly. Thus, the crucial element of the uncertainty monitoring paradigm lies in how these experiments consist of two phases. The first phase is characterized by a forced-choice, so that discriminatory performance will decrease for hard-to-discriminate trials. Crucially, this performance is compared to that observed during the second phase of the experiment, which is characterized by the introduction of a third option for subjects to choose, namely the ‘uncertainty response’ (UR), also called the ‘opt-out’ or ‘skip option’.

Choosing UR removes the present trial, which allows subjects to avoid potential timeouts, but offers no reward, or, in some cases, a less favorable one. Clearly, it is better to opt-out of hard-to-discriminate trials than to suffer the consequences of failing to discriminate correctly, so subjects making strategic use of UR demonstrate a marked increase in performance for hard-to-discriminate trials relative to forced-choice trials in the first phase of the experiment.

The UR-based performance increases of non-human animals are then compared to that of humans, in an effort to determine whether non-human animals use UR as effectively as humans. It is argued that when their results are on par, these animals possess advanced decision-making skills, which suggests to many comparative psychologists that the cognitive systems of these animals possess metacognition (see Proust 2013). Specifically, any animal making strategic use of UR possesses a cognitive system that evaluates its own cognitive capacities and produces feedback signals (noetic feelings) that motivate and guide the animal to act on behalf of those evaluations.

More controversially, metacognition is claimed to be a phylogenetic precursor to capacities for self-knowledge based on three core arguments. First, as discussed above (Chapter 1), metacognitive capacities are argued to enable crucial cognitive features essential to capacities for self-knowledge (see the cognitive role approach; Proust 2013). This is the route pursued here (by both the present case and the alternative one introduced in this chapter), but rather than appealing solely to the role of metacognition in producing a normativity of an epistemic kind, I stress the role of noetic feelings as strongly embodied and appeal to how they enable and facilitate executive control over cognition and behavior, as well as ground both intra-individual and inter-individual regulatory practices of an epistemic kind (see Chapters 4 and 6).
Second, metacognition is argued to be a crucial precursor to capacities for self-knowledge due to how it produces advanced decision-making behavior that plausibly results from the animal’s attentional focus being directed to an internal state, essentially providing the animal with a rudimentary form of introspection (see Smith and Washburn 2005). Though somewhat similar to the present case, this is distinct from the route pursued here (by the present case and the alternative). Rather, in Chapter 4, the mental states characterized by noetic feelings were argued to be proto self-belief states (not proto introspective states).

As a reminder, it was argued above that these mental states ought to be construed as proto self-belief states due to four reasons: (i) mental states characterized by strongly embodied noetic feelings represent (in the de re mode) neurocomputational changes that describe the animal’s own abilities, (ii) these mental states are characterized by conscious cognitive affordances, i.e. they describe functional relationships between internal features and the subject’s abilities, in such a manner that the animal has executive access to these relationships, (iii) these states aim at truth and have mind-to-world direction of fit, and finally (iv) these states enable and facilitate an intra-individual regulatory process (akin to an emotional maturation process) that approximates the revising and updating of beliefs, so that, therefore (i-iv), these states functionally approximate full-blooded states of self-belief. Notice that even if the animal’s attention is directed at external features of the mind-independent world (e.g. directed at bits of the world the animal is engaging with), and not internal features of the animal’s own cognitive processing (i.e. and so does not reflect a case of proto introspection), this would not cast doubt on any of the reasons above that justify construing these mental states as approximating self-belief states.

Moving on to canonical arguments for claiming metacognition is a crucial precursor for capacities for self-knowledge, the third argument will comprise a central concern of the present chapter. Accordingly, metacognition is argued to be a phylogenetic precursor to capacities for self-knowledge due to the pattern with which metacognition is observed across the animal kingdom. For example, Old-World monkeys demonstrate markedly better performance in uncertainty monitoring tests than New-World monkeys, which is precisely what would be predicted if metacognition reflected a crucial precursor to capacities for self-knowledge.

This is because the oldest common ancestor between humans and Old-World monkeys and apes is estimated to have existed 30 million years ago (Ciochan and Fleagle 1987), which is relatively more recent when compared to the oldest common ancestor between humans and New-

---

25 I would like to remind (and flag for the next chapter) that this reason (number iv) is brought to light by the embodied and emotional character of noetic feelings and so we can lose sight of it when noetic feelings are instead conceived of as non-embodied and non-emotional.
World monkeys that existed over 40 million years ago (Moynihan 2015). In other words, Old-World monkeys are more closely related to humans than their New-World cousins, evolutionarily speaking, and so the prospect that Old-World Monkeys have metacognition, while New-World monkey do not, supports claims about metacognition serving as a crucial phylogenetic precursor to distinctly human forms of self-knowledge.

But this line of reasoning is interrupted by the problem of metacognitive biodiversity, and once introduced, the case for metacognition then tends to fall back on the second sort of argument that appeals to proto introspective states. Let consider the details of this debate over the status of metacognition as a phylogenetic ingredient to self-knowledge.

Recent empirical evidence suggests that metacognition emerges in degrees across the animal kingdom, which, depending upon who is asked, serves as either further support for or as a specific kind of counterevidence against metacognition. For example, recent studies with New-World monkeys have demonstrated that these animals will make strategic use of UR so long as the experimental conditions are altered so that the options are increased from the standard “two plus UR” to “six plus UR” (Beran et al. 2016). Proponents of metacognition see this progress made with New-World monkeys as evidence that metacognition comes in degrees, as so often is the case with much of nature’s victories over entropy and disorder (see Smith et al. 2019).

Embracing metacognitive biodiversity, these researchers point to other studies producing similar results. These studies show evidence of weaker forms of metacognition in dogs, where it appears to be dependent on social cues provided by human caretakers (McMahon et al. 2010); in rats, where it appears to be linked primarily with the olfactory sense (Templer et al. 2017); and even progress has been made in demonstrating metacognition in pigeons, specifically how metacognitive capacities are more likely to emerge when pigeons are given visual clues (Iwasaki et al. 2019). One fascinating study discussed in detail below demonstrates evidence of metacognition in honeybees (Perry and Barron 2013; see Section 5).

For skeptics, such progress is evidence that the uncertainty monitoring paradigm exposes not a metacognitive capacity at all, but merely first-order cognitive capacities for monitoring features of the external world and mitigating risk, not monitoring and controlling features of an internal cognitive state, which, in their eyes, disqualifies these capacities as properly metacognitive (Beran et al. 2016; Carruthers and Williams 2019). Notice then, that this debate then circles back around to the second sort of argument, in which the case is made for or against the mental states characteristic of metacognition as introspective (or proto introspective) states.
Here is a brief summary of the skeptic’s reasoning with respect to the recent study on New-World monkeys. So long as only one option is correct, the skeptic points out, decisions made with more options are riskier than those made with fewer options. Thus, these results ought to be explained—not by appealing to introspective states, monitoring and controlling internal features—but by appealing to capacities for monitoring external states of affairs, such as the number of external options to choose from (see Carruthers and Williams 2019).

Although one might push back by pointing out that this interpretation is, at best, specific only to New-World monkeys (and, at worst, ad hoc), this rejoinder fails to shift the conceptual landscape of the debate: it is nonetheless difficult to maintain (or even to envision how one could acquire evidence for the claim) that the advanced decision-making behavior of animals is the result of the animal’s attention being directed to an internal (mental) state, rather than to external features of the world. For this reason, the problem of metacognitive biodiversity needs to be addressed by some other means, specifically by appealing to the first argument, the cognitive role approach.

In sum, it should be clear by now that any adequate theory of metacognition that construes it as enabling both the advanced decision-making behaviors observed in these animals, as well as advanced forms of self-knowledge, incurs a conceptual debt that will need to be repaid in two separate installments. On the one hand, the theory will need to have the conceptual tools to account for this metacognitive biodiversity, while, on the other hand, the theory will also need the tools for making it clear how metacognition can serve as an enabler and facilitator of reflective or executive actions, potentially serving as a crucial precursor to advanced forms of self-knowledge. We shall revisit below how researchers working within the predictive processing framework might address the problem of metacognitive biodiversity while addressing the problem of triviality (Section 5).

Opting-out rather than discriminating during ambiguous trials is the exact behavior comparative psychologists and philosophers believe demonstrates that these animals possess metacognition, that is, mechanisms responsible for producing mental states crucial for the animal acquiring some rudimentary sense of when it knows versus when it does not know (e.g. Beran et al. 2009; Kornell et al. 2007; Washburn et al. 2006; Fujita et al. 2012; Roberts and Lurz 2009). More specifically, this behavior is often argued to be the result of the animal directing its attention to an internal state and acting as a consequence of the content of this state and not the result of the animal directing its attention to features of the environment (see Smith et al. 2019; Templer et al. 2019; Comstock 2019; Hampton 2019; Yuki and Okanoya 2017). Though, as already discussed, this argument is not pursued by the present work due to how it fails to make progress in the debate over metacognition. Before we discuss how to solve the problem of metacognitive biodiversity with the predictive processing framework, let us consider arguments for and against the view that
uncertainty monitoring tasks reveal metacognition in animals by turning to further details surrounding specific cases.

3. Evidence for Animal Metacognition and the Lessons It Teaches

Let us begin by considering perhaps the most famous experiment to have deployed the uncertainty monitoring paradigm that tested an 18-year-old bottlenose dolphin named Natua (Smith et al. 1995). For the most part, this experiment is congruent with previous descriptions, in that it turns upon perceptual discrimination tasks (either high, low, or ambiguous/hard-to-discriminate auditory signals), two phases of the experiment (forced-trials versus UR-trials), and the comparison of animal to human performance in UR-trials over that of forced-trials.

As can readily be guessed by the narrative of our discussion, the results were fascinating: Natua used UR as effectively as any human. In fact, Natua displayed behaviors during difficult trials that readily suggest hesitation much like how humans do. For example, he would swim over to the low paddle, then over to the high paddle, before eventually swimming over and pressing the opt-out paddle, all the while moving his head from side to side and keeping his mouth open.

Several points of interest are found in the details of this experiment, points which are not often discussed and worth going into. First, the conditions under which Natua learned to use UR are quite fascinating in their own right since he had to transfer its use from the original learning setup to the UR-setup. Initially, Natua was trained with the opt-out paddle by dealing with “broken” trials in which neither primary response had any effect and the only way to progress was to press (what would become) the opt-out paddle. Thus, he had to transfer the use of the opt-out paddle from an escape response to the uncertain response, something he did fairly quickly and easily. The ability to transfer the use of the opt-out mechanism from the initial learning condition to a novel setup involving the mitigation of uncertainty is a common approach in comparative psychology for distinguishing this opt-out behavior as actively learned, rather than acquired through associative learning. We will revisit this topic again below.

Also deeply fascinating is that both Natua’s behavior and human behavior differed in the same respect from an ideal observer (a computer simulated subject) that sought to maximize rewards. This fascinating result will be used later on as a means to contextualize and make sense of the problem of the origin of self-knowledge, so it is worth delving into it. As it turns out, instead of opting out of every difficult trial, both human and some non-human animals will tend to opt in
during threshold cases at the expense of reward. Smith et al. explain this somewhat counterproductive behavior by appealing to the complex psychological state of uncertainty that potentially motivates subjects to resolve the uncertainty in an \textit{emotionally satisfying} manner rather than avoid it altogether. This point cannot be overstated. It exposes one central aspect as to why it is so crucial that metacognitive feedback signals become strongly embodied noetic feelings (see \textbf{Chapters 4 and 6}).

Consider again the analogy of playing a videogame. You might know a particular quest or puzzle is extremely difficult and offers little reward in its own right, other than that of having completed the quest and so having bested the game, and so you stubbornly attempt the quest, over and over, until you have finally completed it. Smith et al. believe this somewhat counterproductive behavior suggests something more is going on in the mind of Natua, something which is over and above mere associative, reward-based learning, something quite psychologically complex.

This result is crucial to the present case for shoring up metacognition as a significant precursor to advanced forms of self-knowledge. The central proposal is that the psychological state that explains this behavior is best described as an emotional state, one that is marked by valenced arousal, so that these somewhat counterproductive behaviors are the keys to understanding how metacognition produces mental states that serve as precursors to more advanced states of self-knowledge, those that enable an intra-individual regulatory practice (see \textbf{Chapter 4}), as well as enable inter-individual practices (see \textbf{Chapter 6}).

The psychological need to explore, exploit, make sense of, and ultimately resolve uncertainty is a much-discussed topic within the literature of predictive processing (see Clark 2018; Hohwy 2020). At first glance, behaviors like these pose a problem for the theory: if cognition is in the business of maximizing evidence and reducing uncertainty, why then do subjects seek out and engage the novel, the strange, the unknown and the uncertain, and \textit{enjoy} doing so? Within the predictive processing framework, this is because, as it turns out, routinely engaging with uncertainty is highly advantageous, since it sharpens our skills for coping with it in the long run, making us more effective at handling the uncertainty that could potentially be threatening. The lesson is thus that not all uncertainties are alike. Some kinds of uncertainty are safer, and so apt for exploring, exploiting, and making sense of.

That said, it is an open-ended question whether so-called ‘epistemic foraging behaviors’ are exclusive to highly intelligent creatures like Natua or whether they reflect a more basic fact about how living organisms learn to survive in uncertain environments (see Bruineberg and Rietveld 2014). But since it was demonstrated above that the products of metacognition, strongly embodied
noetic feelings, are likely crucial enablers and facilitators of conscious executive functioning (see Chapter 4), this issue ought not to worry us so much.

Rather, what is deeply intriguing is the prospect that Natua might well have been in a state not all too different from the state humans experience when finding themselves in challenging and uncertain situations, a mental state characterized by a noetic feeling of uncertainty with the need to resolve this feeling in an emotionally satisfying manner so that it can be replaced with a reassuring feeling of confidence. Compare this to the feeling of confidence at having bested the videogame. It feels good in part because it somehow implies you are now better equipped to deal with similar challenges down the road, and if the theories above are correct, this is because it means you have managed to reduce uncertainty in the long run.

Of course, Natua is not the only non-human animal to display advanced decision-making behaviors in the uncertainty monitoring paradigm. Shields et al. (1997) performed similar experiments with two 9-year-old rhesus monkeys, Abel and Baker, who had to categorize a visual stimulus as either dense or sparse in a perceptual discrimination task. In this experiment, both monkeys were able to transfer spontaneously the use of UR from broken trials to trials where UR was optional, and their behavior differed from an ideal observer, much like how human behavior Natua's behavior does. Given our discussion above, the behavior associated with this suboptimal performance (lower than an ideal observer) can be recast as an epistemic foraging behavior, construed as the animal's exploration and exploitation of healthy forms of uncertainty:

“Monkeys, like humans, seem to accept that the escape response is for threshold. In this respect they act more like uncertainty-based machines (using the escape response when they do not know) than like reinforcement-based machines (using the Star response [i.e. UR] to enrich maximally their reward environment)” (ibid: 92).

Though these and similar experiments in comparative psychology were highly influential in demonstrating advanced decision-making skills in non-human animals, skills that are potentially enabled by capacities for monitoring and controlling cognitive operations, not everyone is convinced that these animals possess metacognition, or, let alone, know they do not know. Over the past 25 years, waves of criticism have crashed against these initial experiments and, as a result, new methods have been developed precisely to get at resolving those challenges.
As anticipated by Smith et al. (1999), many skeptics went on to question whether these experiments provide sufficient evidence for claiming the advanced-decision making behavior observed in these animals is not simply acquired through associative learning or operant conditioning, that is, whether animals were not simply using UR to maximize rewards. To that end, novel experiments deployed a technique called ‘deferred feedback’, in which the use of UR was not directly rewarded; instead, rewards/punishments were only doled out after the subject had already completed a block of trials, ensuring UR did not become associated directly with some response (e.g. Couchman et al. 2010).

Other experiments involved prospective judgments, in which subjects had to determine in advance whether to opt-out of trials and subjects were only rewarded or punished at the end of trial-blocks (e.g. Brown et al. 2017). Another approach showed that UR, different from the primary response options, depended upon resources in working memory, thus providing evidence that strategic use of UR depends upon executive, rather than stimulus control (e.g. Smith et al. 2013).

Though there have been some hold-outs (e.g. Le Pelley 2014), these results have been largely successful in convincing skeptics that something more than associative learning is going on in the minds of these animals. Though, whether this all amounts to precursors to self-knowledge continues to be highly controversial. For example, Carruthers and Williams (2019), who continue to raise skepticism about whether metacognition is a precursor to mindreading, had nonetheless this to say recently:

“We should emphasize that the root of objections to the primate-metacognition literature is not that we think the data are better explained in mere associative terms. On the contrary, we fully accept that these (and many other) animals are capable of flexible executively-controlled decision making” (Carruthers and Williams 2019: 279).

Thus, comparative psychologists and philosophers of mind and cognition, unsatisfied with these diverse results, continue to demand evidence that these behaviors are the result of the animals directing attention to internal mental-like states, rather than to external features of the environment, evidence of something not all too different from introspection (see Ferrigno et al. 2019). Meanwhile, skeptics like Carruthers and Williams (2019) continue to argue that these behaviors are the result of first-order appraisals of risk, which do not depend upon, or in any way result from, a metacognitive awareness of uncertainty.
Crucially, this skepticism is not only directed at non-human animals, but also cast on the prospect that humans possess metacognitive awareness of uncertainty as well:

“So, when asked to provide a retrospective explanation of their behavior, [humans] will correctly interpret it as having been guided by uncertainty. That they do so provides no reason to think that they are aware of their own uncertainty at the time they decided to opt out, however. And even if they were aware of it at the time of deciding, it is quite another matter to claim that they chose as they did because of metacognitive awareness of uncertainty rather than because of first-order uncertainty itself. […] We think a better explanation of the data (both macaque and human) is one framed in terms of risk-based decision making” (ibid: 280; emphasis is original).

Notice this criticism is less of a psychological nature, like that above regarding how to establish active learning over mere associative learning; rather, this criticism strikes a somewhat philosophical tone echoing Morgan’s (1903) canon (sort of the Occam’s razor in comparative psychology), that the best explanation of animal behavior is also the simplest.

But of course there is more to this worry than mere methodology. This criticism turns on an unresolved theoretical question, one that is as pressing as it is puzzling, regarding the psychological, cognitive, neurobiological, and physiological underpinnings of confidence and uncertainty, that is, how confidence and uncertainty are realized in brain states, supported by extracerebral bodily states, implemented in cognitive systems and how confidence and uncertainty ultimately become conscious. Meanwhile, lurking in the background, are questions about what it consists in to have metacognitive awareness of uncertainty, rather than a first-order awareness of uncertainty, essentially raising the question of what exactly it means to call something (a process, a capacity, a state, etc.) metacognitive.

Despite a half century of research, it is not immediately clear how best to distinguish between metacognitive and first-order awareness, whether this be of uncertainty or of anything else. As already discussed (Introduction), the most conservative option is construing metacognitive awareness in terms of the deployment of capacities for mindreading (Carruthers 2009b), since it is widely accepted that this cognitive feat depends upon representing mental states in the de dicto mode of representation in which mental representations are represented as mental representations. This construal makes it very clear how such metarepresentational processes are
metacognitive, entailing ‘thinking about thinking’. But this approach presents us with the troublesome issue that non-human animals are excluded right from the outset since, as evidence suggests, capacities for metarepresentation are well beyond their reach (see Krupenye et al. 2017).

What is needed is another way of making sense of metacognitive awareness, one distinct from mindreading, which does not immediately preclude non-human animals while doing justice to the host of comparative studies that show various degrees of advanced decision-making skills across the animal kingdom, skills that are underpinned by cognitive capacities for mitigating uncertainty. But crucially, as Carruthers and Williams are right to demand, this conception of metacognition should aim to justify talk of metacognition as a phylogenetic precursor to self-knowledge or at least make inroads into solving the problem of the origin of self-knowledge.

As has been recently discussed in the literature (see Smith et al. 2019), in order to do arrive at an understanding of the mechanism that would explain the cognitive ingredients for capacities for self-knowledge, we will first need to resolve questions about the cognitive underpinnings of uncertainty and confidence. To that end, I believe the predictive processing framework, together with the Bayesian Brain Hypothesis, can point us in the right direction, that is, so long as several key ideas are kept in mind regarding how best to distinguish strongly embodied noetic feelings from computational signals that approximate Bayesian confidence. Let us turn to this now by considering the problem of honeybee metacognition.

4. The Problem of Honeybee Metacognition and its Lessons for An Account of Metacognition

It is the view of the present work that the possibility of honeybee metacognition needs to be confronted head on. At first glance, the idea that honeybees have metacognition, which a recent study strongly suggests (Perry and Barron 2013), would seem to be problematic for the claim that metacognition is a phylogenetic precursor to self-knowledge. At first glance, if honeybee possess capacities for metacognition, a species whose last common ancestor with humans was over 600 million years ago, then metacognition must be so cognitively inexpensive to implement that it could not possibly reflect an innovation that would lead to advanced forms of self-knowledge.

To refer back to our analogy (see Introduction), the honeybee is so far removed from the chicken, it will be a hard pill to swallow indeed, if the proposal is that the honeybee is the chicken-


like creature that produced the egg that produced the chicken. As far as I know, this specific counterpoint has not been raised, but it does exist in another form whenever it is demanded that an account of metacognition be provided that possesses the conceptual brakes for preventing it from slipping down the slope of triviality (see Introduction and Chapter 3). Thankfully though, we do not have to accept this, since the solution to the problem of triviality offered by the present work (and, as we shall see, offered by Meyniel and colleagues as well) appeals to metacognitive feedback signals that enable and facilitate executive control over cognition and behavior, highly flexible control that, evidence suggests, individual honeybees are not capable of exercising (see Franks et al. 2002; Passino et al. 2008).

That said, there are various ways of interpreting the prospect of honeybee metacognition, and not all of them would spell disaster for the status of metacognition as a crucial precursor to self-knowledge. First, simply going off the fact that the last common ancestor between humans and honeybees was over 600 million years ago does not necessarily lead one to conclude that if honeybees have evaluative metacognition, then so did this last common ancestor (and so did all its descendants). If that were sound, the case for metacognition as a phylogenetic precursor for self-knowledge would be untenable, but I do not suspect anyone would claim this. This is because, assuming honeybees do have metacognition, there is still the possibility that it has a completely independent evolutionary history from that of mammalian or primate metacognition.

On their own, cognitive monitoring and control mechanisms described by metacognition clearly provide powerful survival advantages to any creature who has them. So, it is entirely possible, even plausible, that evolutionary pressures would exist in various ecological niches that would, in turn, demand creatures develop mechanisms for monitoring and controlling cognitive operations in various configurations. The trick is thus explaining how some configurations of metacognition lead to executive control and lay the foundation for mindreading.

Second, if metacognition is indeed a crucial part of the story of how nature made self-knowing animals, it is plausible that metacognition is not an all-or-nothing mechanism, but rather a family of mechanisms with various members, some older, some younger, who are all related to each other to some degree, though often residing in disparate locations, distinct cognitive systems implemented in various forms of cognitive systems, all across the animal kingdom. As introduced above, this is the guiding thought of many comparative psychologists working in the uncertainty paradigm (Section 2).

What is more, if the Bayesian Brain Hypothesis is correct and the theory of predictive processing true, then the mitigation of uncertainty (prediction error/free energy) is at the core of
what brains and their cognitive systems are designed by nature to do, so it should come as no surprise if more basic forms of metacognition exist across the animal kingdom. If these ideas are correct, the prospect of honeybee metacognition actually weighs in favor of metacognition as a phylogenetic precursor to self-knowledge, albeit the kind found in honeybees ought not to be construed as a crucial precursor in the exact sense articulated above (Introduction): it is not as though honeybee metacognition lays the foundation for the origin of self-knowledge (i.e. in our go-to analogy, it is not the chicken-like creature that produced the egg that produced the chicken).

Alternatively, honeybee metacognition ought to be construed as a very early precursor (i.e. a chicken-like creature that produced an egg that produced another chicken-like creature, but not the chicken, and on and on, however many iterations back), a construal that demands two accounts be adequately provided. On the one hand, some account is required of how evaluative metacognition evolved in honeybees, or, at the very least, how evaluative metacognition can be implemented in the cognitive systems of honeybees. I think the resources to develop a plausible sketch of this are available and this will be discussed below.

On the other hand, in order to hold fast to the idea that metacognition plays a crucial role in the emergence of self-knowledge, some account of the mechanisms underpinning and enabling the advanced decision-making skills in primates needs to be provided that allows us to clearly distinguish it from simpler mechanisms like that found in honeybees. In particular, it must be clear why the former mechanism (that found in mammals) reflects a precursor to capacities for self-knowledge while the latter (that found in insects) does not. But before delving into all that, let us first examine the evidence of honeybee metacognition in detail.

Perry and Barron (2013) devised an experiment based on the uncertainty monitoring paradigm to test whether honeybees can opt-out of difficult choices so as to improve their performance, an experiment that included testing whether bees could transfer their uncertainty-mitigating skills to novel tasks. Free-flying bees entered into a two-chamber apparatus. The first chamber presented bees with a discrimination task about whether a visual target was above or below a reference bar. Approximately, half the bees were rewarded for correctly discriminating the above relation, while the other half were rewarded for correctly discriminating the below relation. These tasks could be made more difficult, and even impossible, by positioning the target image closer to the exact center of the reference bar.

Meanwhile, the first chamber had a small hole serving as an exit port which simultaneously allowed the bee to opt-out of tasks, as well as progressing with the experiment, by travelling to the second decision chamber. This chamber lacked an exit port and so provided the forced-choice task
with which to compare discrimination performance. The results were indeed fascinating. “Bees opted out more often on hard or impossible trials than on easy tasks” and when their performance was compared between unforced and forced hard trials, bees performed better on unforced hard trials than on forced hard trials, “indicating that bees opted out selectively when their likelihood of successfully solving the task was lowest” (ibid: 19157).

Skill transfer, a key approach to pushing back against the reductionist strategy of reducing these behaviors to mere associative learning (see Section 2 above), was determined in one of two distinct ways. First, bees were presented with novel visual stimuli and had to transfer the skill of mitigating uncertainty to these novel situations. This task presented little difficulty for our little friends. Second, and much more challenging, bees were presented with confusing targets for which the task did not actually make sense, since the reference bar was entirely absent. Here, it was always in the bee’s best interest to always opt out. However, these results were somewhat mixed: “no bees opted out during training [with unrewarded confusing targets], suggesting they did not select to opt out in response to a novel stimulus” (ibidem). That said, “4 of 10 bees opted out […] on their first trial […],” that is, after training, “and at least three of the five test trials with the confusing targets” (ibidem). That’s not too terribly much, but it is something.

Though these results provide some evidence of skill transfer in honeybee metacognition and so ward off reductionist strategies construing these behaviors as the consequence of mere associative learning, these results contrast with observations in Old-World monkeys who demonstrate robust forms of evaluative metacognition. Kornell et al. (2007) showed that both rhesus monkeys, Lashley and Ebbinghaus, were able to transfer quickly and easily their metacognitive skills for mitigating uncertainty from a perceptual to a memory task. This difference in cognitive flexibility between primate metacognition and honeybee metacognition evidences the proposal that evaluative metacognition comes in degrees, that crucial behavioral differences are underlaid by distinct cognitive capacities (more on this in Chapter 6).

One final important point concerns whether honeybee metacognitive performance differs from that of an ideal observer. Recall that Old-World monkeys, cetaceans, and humans all perform worse than an ideal (Bayesian) observer who sought to maximize rewards. Recall further that this is regarded as evidence of there being a psychologically rich emotional state at play in this behavior, which is interpreted by the present work to suggest these states demand noetic feelings of uncertainty be resolved into emotionally satisfying noetic feelings of confidence (see Chapter 6). That said, it is not known, as it was not controlled for, whether the metacognitive performance of honeybees differs from an ideal observer in similar ways that would suggest their behavior is motivated by emotional episodes.
Thus, when the performance of honeybees is contrasted with the performance of mammals, a sensible proposal results. Though honeybees might possess representational systems for producing cognitive affordances, it is unlikely, however, that these affordances involve conscious forms of executive control. If they were conscious forms of executive control, one would expect to observe better performance in tasks of skill transfer. What is more, it is unlikely that these affordances are conscious expressions of emotion, noetic feelings. If affordances are not conscious, it follows then that they are also not conscious expressions of emotion.

While claims about executive function, capacities for skill transfer, and consciousness in honeybees invite further debate, conclusive cases for either topic need not provided here (for a discussion of this, see Eisenreich et al. 2017). The challenge of the present discussion is to provide a plausible way of thinking about honeybee metacognition that avoids conceptualizing human metacognition as trivial in terms of how it might serve as a crucial precursor to self-knowledge.

Thus, the conclusion is that honeybee metacognition ought not to be regarded as a crucial precursor to self-knowledge, but a very distant precursor. To echo the formulation above (Introduction), honeybee metacognition was not the chicken-like creature that made the chicken-egg that made the chicken. Rather, it was, at best, this creature’s distant relative. This is because there is inadequate evidence to suggest that the cognitive control mechanisms of honeybee metacognition enable conscious executive control of behavior (let alone, evidence this serves as a significant precursor to self-knowledge). Were this the case, there would be evidence of multimodal integration and domain-general application observed in cases of executive control.

With all of this being on the table, let us now discuss an alternative proposal to the one offered by the present work, a prominent model for integrating metacognition within the predictive processing framework. As shall be discussed, it too will be faced with solving the problem of triviality, its downstream problems of metacognitive biodiversity and honeybee metacognition, as well as the ultimate problem on the origin of self-knowledge. Whereas it solves the first three in a similar manner as the present work, it will be shown to struggles to account for the ultimate problem, details of which will be discussed in the final chapter (Chapter 6).
5. Distributional Confidence on the Path to Subjective Confidence: almost home, but not quite yet.

In this section, we will discuss a prominent model for integrating metacognition into the predictive processing framework. To begin, consider how Perry and Barron (2013) explain the advanced decision-making behaviors observed in their experiment as possibly resulting from a process that enables the “direct assessment of uncertainty” (ibid: 19158), the underlying mechanism of which Perry and Barron describe within a neurocomputational framework. It begins with an ambiguous stimulus that coactivates two conflicting motor responses and whose resolution is achieved through mutual inhibition, “so that only a single dominant motor pattern is activated” (ibidem). For making sense of the details, Perry and Barron appeal to two, what have become, highly influential sources in the literature on predictive processing, one that describes the neurobiological level, and another for the cognitive level of description.

Concerning the neurobiological underpinnings of uncertainty-based decision making, Yu and Dayan (2005) discuss evidence that these behaviors are influenced by the neuromodulators acetylcholine and noradrenaline, one for expected uncertainty and another for unexpected uncertainty, respectively (for the topic of neuromodulators in the honeybee brain, see Galizia et al 2011). As it regards the cognitive-computational level of description, Perry and Barron appeal to Daw et al.’s (2005) seminal work that views decision-making behavior from the vantage of the systems theory of cognition introduced above (see Introduction). These researchers model decisions also in a manner congruent with the Bayesian Brain Hypothesis and predictive processing. Meanwhile, both models (Yu and Dayan’s and Daw and colleagues’) enjoy wide acceptance in the literature (see Schwartenbeck et al 2019; Hutchinson and Barret 2019; Kuperberg et al 2020; Inglis et al 2021).

Appealing to these same models, Meyniel et al. (2015a; 2015b) offer an explanation of confidence signals that spans the various domains of the brain sciences, from neuroscience to psychology, culminating in a view that enjoys strong influence in the literature on predictive processing (see Rollwage and Fleming 2021; Legrand et al 2020; Suárez-Pinilla et al. 2018). As a result of this breadth, their account has the potential to illuminate how subpersonal Bayesian confidence signals can transform into psychological states characterized by noetic feelings that motivate and guide advanced decision-making behaviors observed in experiments based on the uncertainty monitoring paradigm, and so can potentially help us make sense of any cognitive differences between honeybee and mammal metacognition.
Meyniel and colleagues identify two forms of confidence: “a distributional form of confidence that pertains to probabilistic representations and a summary form that pertains to scalar representations derived from those distributions” (2015b: 78; original emphasis). The former refers to the neurocomputational level of description and the latter describes the psychological level, even though Bayesian probability remains “the formal definition of subjective confidence” (ibidem). The difference between these forms of confidence is analogous to the difference between summary statistics and the data sets described by them.

Consider the notion of standard deviation, routinely used for calculating statistical relevance. This is a measure of the so-called ‘spread of data’. If deviation is low, data are clustered around the mean. If deviation is high, data are dispersed over a range of values. Crucially, standard deviation is a scalar-based summary that describes a particular set of data. So think of distributional confidence as the Gaussian spread of data, while summary confidence is the scalar value describing standard deviation from the mean of the spread.

Starting at the neuro-computational level of description, Meyniel et al. account for the distributional form of confidence by appealing to the Bayesian Brain Hypothesis, which maintains that activity in neural populations are representations approximating Bayesian probability distributions. Probability distributions are subpersonal representations that are defined solely by their computational role (see representations of the visual system discussed by Marr 1982). Crucially, this form of representation does not need to become accessible at the personal level.

Meyniel et al. propose thinking about these representations as passed through specific readout mechanisms that perform linear integrations and yield scalar representations of the distributions. One such readout mechanism is for yielding summary confidence. Though subjects rely upon summary confidence when asked to provide ratings of subjective confidence in various psychological experiments, such as in the uncertainty monitoring paradigm, the readout mechanism that produces summary confidence is crucially distinct from the mechanism that determines choice and enables sensori-motor behavior.

This is because, while scores of operations rely upon distributional confidence (see TOTE units, reinforcement learning, reafferent feedback, etc.), only a very exclusive set of operations utilize summary confidence. Only those operations that are responsible for driving the behavior of the entire animal, and not merely driving one of its subsystems, are theorized to utilize summary confidence. As such, mechanisms that produce summary confidence are distinct from, but also crucially intertwined with, mechanisms that enable executive control. In what follows, we will discuss the detail of such readout mechanisms, how they relates to the conscious control of
Highly relevant for our discussion, Meyniel et al. analyze the neuro-computational underpinnings of opt-out behavior observed in the uncertainty monitoring paradigm. Choices like these, they argue, require confidence be “derived as accurately as possible from the subject’s internal representation of that information” (2015b: 83). That said, Meyniel et al. do not think summary confidence is equivalent to subjective confidence. This is because studies on subliminal reinforcement learning have shown that subjects become “confident in the reward delivery systems, which is demonstrated by their choices and the neural prediction errors observed in case of violations, but they remain unaware of it” (ibid: 86). By which they mean, subjects are unaware of the representations of summary confidence that underlie their confidence-based behavior (see Pessiglione et al. 2008).

Pessiglione et al. (2008) demonstrated the effectiveness of masked contextual cues in instrumental learning. When presented with masked cues that suggest trial difficulty, subjects perform discrimination tasks and wager about likelihood of success. Though their behavior suggests their decisions are guided by these masked cues, subjects are unable to report on them or even discriminate between them when later unmasked, expressing surprise upon seeing them. Pessiglione and colleagues interpret this as evidence that intuitive decision-making need not depend upon conscious information, since, as their study demonstrates, these decisions are guided by subliminal processing.

Thus, summary confidence is theorized to be operative in driving the behavior of the whole animal, while not necessarily operative at the level the animal has access to, that is, the level where it makes sense to talk about executive control. But subjective confidence is different. Subjective confidence is theorized to be conscious and accessible to the executive control network. Recall from the previous chapter (Chapter 4), how an appeal to executive control comprised an essential part of solving the problem of triviality facing metacognition. Here, likewise, an appeal to subjective confidence will help shore up Meyniel et al.’s approach to integrating metacognition within the predictive processing framework while avoiding the various problems discussed above.

Meyniel et al. provide a sketch of the process through which summary confidence could become consciously accessible subjective confidence. From a neuroanatomical perspective, they suggest adopting a highly nuanced view, in which different neural circuits contribute to the production of confidence depending on behavioral needs. In this respect, they recommend investigating both cortical regions, particularly prefrontal areas, as well as subcortical regions, such
as the ventral striatum that serves as an interface between the motor system and the limbic system (see Groenewegen and Trimble 2007). We will revisit the limbic system’s possible implication in the production of subjective confidence in the next chapter (Chapter 6). There it shall be argued that the limbic system is crucial to shoring up metacognition as a significant precursor to advanced forms of self-knowledge, as it essential to producing strongly embodied noetic feelings.

At the level of cognitive-psychological description, Meyniel et al. suggest appealing to the global workspace theory (Baars 2005) to account for how summary confidence could become subjective confidence. Meanwhile, Hohwy (2013) also proposes adopting the global workspace theory within the predictive processing framework. Essentially, he explains the global workspace by appealing to representations that serve as top-down and laterally-flowing generative models of incoming bottom-up sensory data that compete with one another over providing the most accurate description of these data. Once the brain is in the position to select a policy for action, the winner of this competition becomes broadcast globally and available for conscious report.

This is the so-called process of ‘ignition’, which Hohwy envisions as involving the process of prediction error minimization, in which generative models acquire evidence until they warrant guiding action. Effectively, Meyniel et al. suggest distributional confidence goes through the process of ignition, becoming ‘read-out’ by a cortical-subcortical neural circuit, before transforming into a globally broadcast, conscious signal of confidence. Note, however, this conscious signal of confidence is not identical to a strongly embodied noetic feeling of confidence, since the neurocomputational changes described by this signal are not theorized as being integrated within the executive control network (and so conscious) through the interoception of bodily afferents (see Chapters 2, 4, and 6).

Thus, on the table is an alternative proposal to the one offered by the present work for addressing the problems facing metacognition. Whereas the present work appeals to strongly embodied noetic feelings, Meyniel et al. provide an account that appeals to non-embodied (central and solely cerebral) conscious metacognitive feedback signals. Moreover, their account consists in distinguishing subjective confidence from summary confidence, and thus distinguishing those representations of distributional confidence that become available to the global workspace, available at the animal-level of description and so conscious.

This proposal is likely to be successful given the discussion in previous chapter about the need to appeal to the executive control network and conscious access to shore up metacognition against the problem of triviality. Essentially, the proposal amounts to appealing to consciousness to solve the four-fold problems above (the problem of triviality, metacognition biodiversity,
honeybee metacognition and (possibly) the origin of self-knowledge), so let us examine this proposal in more detail.

First, one would need to wager that success in uncertainty monitoring tasks is due to certain cognitive advantages obtained by the utilization of summary confidence by mechanisms that control whole animal behavior.

“The opt-out example illustrates how choices based on uncertain information can serve to measure a subject’s summary confidence. Insofar as people and animals seek to optimize their gains, these choices require that summary confidence is derived as accurately as possible from the subject’s internal representation of that information. That is, optimal wagering decisions require optimal readout of summary confidence from distributional confidence, provided that such information is available to inform the outcome.” (Meyniel et al. 2015a: 83)

According to this proposal then, the problem of metacognitive biodiversity could be explained by arguing that, for example, in the case of dogs, the mechanism that explains their interactions with human caretakers is perhaps one of the few that can utilize summary confidence. As such, there is a domain-specific readout mechanism in dogs that is wired through those neuroanatomic structures responsible for the dog’s social cognition, in particular that subpart that underlays its social cognition with humans, and this mechanism is one that crucially utilizes a scalar representation of distributional confidence.

Meanwhile, a similar story could be used to explain the differential sets of skills in the other animals. For New-World monkeys, the crucial readout mechanism is domain-specific to conflict resolution and risk mitigation realized in neuroanatomical structures, such as the cingulate cortex (see Teuchies et al. 2016; Paus 2001). For rats, the mechanism is tied to the olfactory bulb. For pigeons, the mechanism is tightly bound to the visual system.

However, since modest success in tasks of skill-transfer were demonstrated by honeybees, their cognitive control mechanisms may or may not be domain-general. Empirical evidence for this could be gathered either way by investigating whether honeybees are able to transfer their skills for mitigating uncertainty quickly and easily from the visual to the olfactory domain. Therefore, as it currently stands, we do not have enough empirical evidence to rule out the possibility that honeybees have domain-general metacognitive capacities.
In summation, to solve the problem of metacognitive biodiversity and account for the borderline cases of metacognition in some non-human animals, one can appeal to domain-specific signals that drive the animal’s behavior, signals which are not produced by the global ignition process, nor integral to executive functioning (i.e. highly-flexible, domain-general cognition and behavior). Within Meyniel et al.’s account this essentially means, though the behavior of these animals is driven by representations of summary confidence, the mechanisms underpinning metacognition in these animals reflects only modest forms of cognitive sophistication compared to basic feats of cognition, since it is unlikely that these mechanisms enable capacities for executive control over cognition and behavior made possible by subjective confidence.

Appealing to the difference between domain-general (and so conscious and executive) and domain-specific signals (and so unconscious and stimulus driven) might succeed in explaining how this evaluative process is indeed a sophisticated one (see Morales et al. 2018), and so Meyniel et al.’s theory can ward off attacks that would reduce metacognition to the day-to-day business of cognition, thus succeeding in solving the problem of triviality. Similar to the solution appealed to above by the embodied metacognition approach (see Chapter 4), Meyniel et al.’s approach depends on an appeal to conscious executive control over cognition and behavior.

That said, there is a significant issue awaiting this proposal. As already discussed, recent empirical evidence from interoception-based studies on metacognition demonstrates the degree to which metacognitive feedback signals, noetic feelings, are strongly embodied (see Chapter 2). In particular, this evidence makes plausible a Neo-Jamesian theory about noetic feelings, such that their role in executive functions is enabled by the interoception of bodily afferents. This issue (how it is problematic and how it can be addressed) will be discussed below (Chapter 6).

What is more, it is unclear how even domain-general metacognitive signals are supposed to explain the emergence of self-knowledge. That is, while it is clear that it succeeds in solving the problem of triviality (and its related problems of metacognitive biodiversity and honeybee metacognition), it is nonetheless unclear how an appeal to conscious feedback signals that can be utilized across the various domains of perception, memory, and decision-making can succeed in solving the ultimate problem on the origin of self-knowledge.

Essentially, this is because of how these remain cases of first-order cognition and so do not imply de dicto mode of reference or mindreading capacities. Recall from above, unless we have some reason to believe these metacognitive signals are conceptual representations or characterize full-blood states of belief, then there is nothing to indicate a metarepresentational construal. Otherwise, one could argue that these metacognitive signals characterize proto states of self-belief and due to
their role in cognition and behavior ought to be construed as crucial precursors to self-knowledge. But this is precisely the approach taken below, except it will be argued by appealing to how strongly embodied noetic feelings are better suited due to their status as strongly embodied (Chapter 6).

In the last chapter, the problem of the origin of self-knowledge will finally be confronted. It will be argued there that, though much research of both an empirical and conceptual nature needs to be done, embodied metacognition is better suited than Meyniel et al.’s account to address this problem. To get a glimpse of this, contrast the two proposals. While Meyniel et al. appeal to conscious feedback signals that enable and facilitate executive action, these signals are theorized as solely central and cerebral, whose production does not in any way depend upon the extracerebral body. As such, Meyniel et al.’s account can readily be categorized as a traditional approach argued above to encounter problems when accounting for the mechanisms of metacognition (Chapter 2). Thus, the details of this will be discussed in the next chapter with the backdrop of addressing the problem of the origin of self-knowledge.

6. Conclusion

In the chapter, we discussed the close conceptual proximity that metacognition has to the theory of predictive processing, so close that metacognition, as it is often construed, would seem to be little more than the day-to-day business of cognition, inviting the problem of triviality, and potentially robbing metacognition of its status as a crucial phylogenetic ingredient in the origin story of self-knowledge.

We discussed an approach by Meyniel and colleagues for shoring up metacognition within the predictive processing framework. These researchers appeal to the global workspace theory to discriminate conscious metacognitive confidence signals from those signals that are part of the daily business of cognition. As such, the crucial signals drive executive level control over cognition and behavior. It was argued that this appeal to conscious signals can solve the problem of triviality, as well as its related problems of metacognitive biodiversity and honeybee metacognition. However, it was also argued that this account does not explain evidence from interoception-based studies on metacognition, nor does it yet provide the necessary tools for building a bridge between the phylogenetic and ontogenetic stories about the origin of self-knowledge.
Meyniel and colleague’s appeal was shown to be conceptually quite similar to the appeal made by the present case for embodied metacognition, in that both point to conscious signals that enable executive control over cognition and behavior to solve the problem of triviality and related problems. The difference between the two lies in how the former is a traditional view, while the latter is a strong embodied view of metacognition. In other words, Meyniel and colleague’s account does not appeal to any extracerebral process to explain how metacognitive signals are produced, nor how these signals come to play their role in enabling and facilitating conscious control over cognition and behavior.

In the next chapter, I shall discuss how Meyniel et al.’s traditional approach can be augmented within the predictive processing framework to become a strong embodied account. Thereafter, I shall discuss reasons for preferring an augmented account over an unaugmented account. To that end, there will be ‘backward-looking’ reasons that emerge from the discussion in earlier chapters (Chapters 2 and 4). Meanwhile, with respect to offering the tools for building a plausible bridge from the phylogenetic to the ontogenetic origin of self-knowledge, there will also be ‘forward-looking’ reasons that point to testable empirical predictions, which shall be discussed at the end of the present work should an embodied approach be endorsed (see Conclusions).
Chapter 6

How Feeling Our Hearts Enables Us to Feel Our Minds: Uniting the Phylogenetic and Ontogenetic Stories about the Origin of Self-knowledge

Abstract

In the previous chapter (Chapter 5), I discussed an alternative to the one offered by the present work for shoring up metacognition against the problem of triviality put forward by Meyniel and colleagues. There it was demonstrated how this approach shares a conceptual similarity to approaches discussed earlier (Chapter 2), namely non-embodied, so-called ‘traditional’ approaches, and it was argued in this earlier chapter that traditional approaches run counter to emerging empirical evidence from interoception-based studies on metacognition that demonstrate the essential role that the extracerebral body plays in the overarching process of conscious control over cognition and behavior. What is potentially more, it was discussed how the alternative approach leaves questions unanswered about how metacognition could serve as a crucial phylogenetic ingredient in the origin of self-knowledge.

In what follows, I discuss how Meyniel and colleague’s account can be easily augmented to describe an embodied account of metacognition within the predictive processing framework by appealing to Chanes and Barrett’s (2016) account of the limbic workspace. Thereafter, I argue that the augmented, embodied account ought to be preferred due to reasons already discussed in earlier chapters (see Chapters 2 and 4). What is potentially more, I argue that the embodied account of metacognition, with its strongly embodied noetic feelings, offers the required tools for building a bridge between the phylogenetic and the ontogenetic origin stories about self-knowledge. In particular, this is due to how strongly embodied noetic feelings enable and facilitate both intraindividual and interindividual regulatory practices of an epistemic kind that could serve as the foundation upon which a collective epistemic agency is built, the source from which capacities for acquiring self-knowledge spring.
1. Introduction

As discussed above (Introduction), metacognition is often regarded as a phylogenetic precursor to cognitive capacities that enable the acquisition of self-knowledge. Recall how skeptics question whether metacognition, with its close conceptual proximity to cybernetic theories of non-reflective, first-order cognition, is sufficiently robust to prevent conceptualizing basic feats of cognition as metacognitive, feats which have little to do with reflective or executive processes. In Chapter 4, this ‘problem of triviality’ was solved by appealing to strongly embodied noetic feelings that enable and facilitate conscious control over cognition and behavior.

Meanwhile, in Chapter 5, Meyniel and colleagues’ (2015a, 2015b) account of confidence was demonstrated to offer an alternative to the embodied metacognition approach, though it was interpreted as providing an analogous solution to the problem of triviality. Namely, Meyniel et al. appeal to subjective confidence, conscious feedback signals that enable and facilitate executive control over cognition and behavior. However, it was remarked how this approach is conceptually similar to traditional approaches discussed earlier (Chapter 2), in that in explaining conscious confidence and uncertainty signals, it does not appeal to the extracerebral body. Below I shall argue that its status as a traditional account has the unfortunate consequence that it runs counter to emerging empirical evidence (Section 4). Moreover, Meyniel et al.’s account leaves questions unresolved about how best to go about addressing the ultimate problem facing evaluative metacognition, the problem of the origin of self-knowledge.

Recall from the Lay Summary and Introduction that the problem of the origin of self-knowledge, as it is interpreted here, essentially involves bridging two distinct accounts from cognitive psychology about the origin of self-knowledge. On the one hand, the social-scaffolding view in developmental psychology explains the ontogenetic origin of self-knowledge as found in the enculturation process during a child’s development. On the other hand, comparative psychology explains the phylogenetic origin of self-knowledge as ultimately the result of mechanisms for evaluative metacognition, an account which has been augmented by the present work to become embodied metacognition, in which strongly embodied noetic feelings play an enabling and facilitating role in conscious control over cognition and behavior (see Chapter 4).

As of yet, it is unclear how an appeal to conscious signals, whether solely cerebral, as they are in Meyniel et al.’s view, or strongly embodied, as they are in the present account, can bridge these two theories about the origin of self-knowledge. Thus, this chapter aims to make inroads into solving this problem by introducing the plausible tools needed to build this bridge, and it will be
argued that the embodied metacognition approach of the present work ought to be preferred to the traditional cognition approach offered by Meyniel et al. due to, in large part, strongly embodied noetic feelings. Specifically, embodied metacognition is to be preferred over Meyniel et al.’s non-embodied account due to, on the one hand, evidence discussed above demonstrating the embodiment of metacognition (see Chapter 2 and 4), as well as, on the other hand, considerations about which particular features are required to build a plausible bridge between the ontogenetic and the phylogenetic theories on the origin of self-knowledge (Section 4).

We shall begin below by discussing once again Meyniel et al.’s approach by focusing specifically on how it accounts for conscious metacognitive feedback signals (Section 2). But rather than rejecting this approach outright, this chapter aims to augment it by incorporating it into a prominent theory within the predictive processing framework about the neurocognitive mechanisms of emotional experience (see Barrett 2017; Section 3). Thus, the first goal of this chapter is to demonstrate how Meyniel et al.’s traditional approach can be easily transformed into an embodied one. As such, the present work ought to be seen as a complement to this research.

However, the overarching aim of this final chapter is to provide a sketch of a plausible solution to the problem of the origin of self-knowledge. Instead of arguing, as others have attempted to do previously, that mental states characterized by noetic feelings ought to be construed as precursors to states of self-knowledge due to how they serve as rudimentary introspective states (see Smith et al. 2019), an approach that continues to hold no sway with skeptics (Carruthers and Williams 2019), the present work pursues an argument with three prongs, the first prong having already been discussed above (Chapter 4). Ultimately, mental states characterized by strongly embodied noetic feelings will be argued to be crucial precursors to states of self-knowledge due to how these mental states plausibly play a crucial role in producing the socio-cultural practices that aim at regulating belief formation.

In other words (see Introduction), I shall argue that embodied metacognition is the chicken-like creature that produced the egg (the socio-cultural practices) that produced the chicken (self-knowledge). Thus, mental states characterized by strongly embodied noetic feelings ought to be construed as crucial to the emergence of the socio-cultural practices that aim at regulating belief formation due to how these states (i) index proto self-belief states of self-confidence (already argued for above; see Chapter 4), (ii) tend to promote rich epistemic foraging behaviors in the subject (also argued for in Chapter 4 but expanded upon in Section 4 below), and (iii) tend to possess corresponding external bodily expressions that can be detected, monitored, and regulated by other individuals in the subject’s socio-cultural environment to ground inter-individual regulatory practices of an epistemic kind (Section 4). At the end of the present work
(Conclusions), I shall introduce future lines of research and several empirical predictions that arise from this proposal. Before we discuss the details of this proposal, let’s review key details concerning Meyniel and colleagues approach.

2. Review of the Hierarchical Theory of Confidence: from distributional confidence onto feelings of confidence

Before it can be demonstrated how Meyniel et al.’s account can be modified to become an embodied account of evaluative metacognition, let us review how this account theorizes confidence (and its inverse, uncertainty) to work at various levels of the cognitive hierarchy. Based on the Bayesian Brain Hypothesis (Knill and Pouget 2004) and the theory of predictive processing (Hohwy 2013; Clark 2016), Meyniel et al.’s (2015; 2015b) account of metacognition can be construed as distinguishing four explanatorily distinct kinds of confidence. Starting with the most basic and moving forward with progressively more cognitively complex kinds, there is (i) distributional confidence, (ii) local summary confidence, (iii) wide summary confidence, and (iv) subjective confidence. The crux of the present argument is that Meyniel et al.’s account of metacognition can be readily modified to describe a fifth kind, which it does not yet describe, namely (v) a strongly embodied noetic feeling of confidence.

At its most cognitively basic, confidence is (i) distributional confidence, a pattern of neurocomputational activity, such as the distribution of firing rates across neurons, that represents a probability distribution, the upshot of an inferential process approximating Bayesian inference (see Meyniel et al. 2015: 80). Distributional confidence is further theorized to serve as input to various read-out mechanisms, which perform a linear summation on distributional confidence to yield a summary statistic for the purpose of ultimately influencing action selection (ibidem). This scalar representation is thus (ii) summary confidence, which is essentially an estimation of the precision of the probability distribution (ibid: 82).

Recall from above (Chapter 4) that precision estimations can be construed, for present purposes, as estimations of the priority that certain computations possess within their dedicated control structures compared to that of other computations, either in the same control structure or within other structures in the overarching cognitive system. Within Meyniel et al.’s account,

---

26 Since it has been the topic of previous discussion, I should flag that this notion of ‘represents’ is that of indexing (or something deeply similar to indexing) in the sense described above (see Introduction, Chapters 1 and 3).
summary confidence can thus be construed as describing the computational priority assigned to specific patterns of neurological activity: high summary confidence means high priority, while low summary confidence means low priority, with respect to the influence that the corresponding representations will exert on downstream processing, possibly even influencing those levels that subserve executive control over cognition and behavior.

Let’s now consider the difference between (ii) and (iii), ‘local’ versus ‘wide’. While both forms of summary confidence are conceived of as functioning analogously to feedback signals in cognitive control mechanisms, thus informing about the adequacy of cognitive operations in meeting specific control objectives, the difference lies in the respective degree of influence on downstream processing. In other words, summary confidence is theorized as exerting either a local influence so that its reach might only encompass relatively basic and covert cognitive operations (e.g. those involved in determining the reward value of a stimulus), or a wide influence so that the feedback signals might guide actions involving the whole organism, e.g. decision-making behaviors such as discriminating the denser of two visual stimuli.

The key difference between (ii) and (iii) therefore lies not in the signals themselves but in the configuration of the cognitive control mechanism that produces them, which Meyniel et al. refer to as a ‘readout’ mechanism. If the readout mechanism has only a local reach, the signal will exert only a local influence, but if the readout mechanism is more extensive, tied to more processing centers, the signal will have a wider impact. Below Meyniel et al.’s account will be augmented to describe a readout mechanism for producing strongly embodied noetic feelings, thus modifying the account to be a theory about embodied metacognition.

Notice that it would be bit of a misnomer to refer to wide summary confidence as ‘global’. This is because wide summary confidence is not necessary globally broadcast in the sense of becoming accessible for conscious report (ibid: 86). Rather, summary confidence is conceived of as used for driving whole organism behavior, even decision-making behavior, such as making choices in a gambling task (see Pessiglione et al. 2008), though the subject need not be conscious of those signals driving these decisions. Thus, wide summary confidence is theorized to be explanatorily distinct from a conscious kind of confidence, which comprises the fourth (and final) kind of confidence appealed to by Meyniel et al., namely (iv) ‘subjective confidence’.

Subjective confidence is thus the upshot of summary confidence (either (ii) or (iii)) once accessed by global processing: “Global processing may nevertheless be able to access summary forms of confidence, as, for instance, the level of accumulated evidence can be monitored” (ibid: 86). Here, Meyniel et al. appeal to Dehaene’s (2001) theory of the Global Neuronal Workspace.
(GNW) likewise based on Bayesian models of neural networks (see Pasquali et al. 2010) that describe high-order cortices accumulating evidence about the state of other cortices, a process by which “incoming evidence is accumulated and amplified non-linearly into a full-blown state of high-level activity” referred to as ‘global ignition’ (Dehaene et al. 2014: 5).

Accordingly, to transform into subjective confidence, summary confidence is theorized to undergo the process of global ignition, in which prediction errors are minimized until enough evidence has been acquired to facilitate executive action selection (see Hohwy 2013). All of this implies that subjective confidence is thus the product of three explanatorily distinct waves of error reduction: an initial wave involved in determining distributional confidence in the lower-order cortex, a second wave for determining the summary confidence that the higher-order cortex produces while monitoring and evaluating the lower-order cortex, and finally a third wave for broadcasting confidence globally. Thus, Meyniel et al.’s theory of subjective confidence (i.e. conscious confidence) is ultimately a kind of confidence-cubed.

One final aspect of Meyniel et al.’s account of the process by which summary confidence becomes subjective confidence is that the third wave of error reduction receives a specific neuroanatomical description, so that it crucially involves higher-order cortices accumulating evidence about the state of other cortices. Thus, the global ignition process is theorized to depend upon prediction error minimization occurring across specific neuroanatomic regions, often labelled the ‘rich club network’ so that “conscious processing corresponds to a massive cortico-cortical exchange of information” (Dehaene et al. 2014: 8).

In the next section, it will be argued that the limbic system is the relevant higher-order cortex operating over the other cortices and playing a crucial role in the production of a strongly embodied noetic feeling of confidence. This is because of the limbic system’s laminar structure, which suggests it likely rests atop the cortical hierarchy, constraining and shaping the predictions of the other levels, as well as, of course, the pivotal role the limbic system plays in emotional processing (see Chanes and Barret 2016).

In summation, Meyniel et al.’s account has the conceptual tools needed to solve the problem of triviality, specifically it can appeal to the mechanisms that produce subjective confidence. But it is nonetheless unclear whether this account can properly address the problem of the origin of self-knowledge. To do so, it would need to make it clear why metacognition ought to be considered a precursor to metarepresentational metacognition and self-directed mindreading. It is not enough to beautify the story of metacognition so as to make it more cognitively precious, involving higher-order networks and conscious signals. One must also show that these upgrades
ultimately purchase a crucial ingredient to self-knowledge. But, as already discussed (Chapter 5), it remains unaddressed and unclear exactly how conscious confidence signals play a crucial role in explaining how subjects come to know that they know.

That said, Meyniel et al. have some options available that will be discussed in the final section of this chapter. For example, they might appeal to external manifestations of subjective confidence in order to build a bridge from the phylogenetic to the ontogenetic origin of self-knowledge. Though, for reasons that will become clear, the embodied metacognition approach ought to be preferred due to evidence discussed above that leans in favor of an embodied mechanism for producing conscious confidence signals (see Chapters 2 and 4), as well as, how an embodied approach has more tools at its disposal to build this bridge and thus make plausible inroads into solving the problem of the origin of self-knowledge.

Regarding recent empirical evidence that supports an embodied view, it will be argued below that Meyniel and colleagues’ unaugmented account fails to account for this evidence. What is potentially more, it will also be argued below (Section 4) that it is plausible that the problem of the origin of self-knowledge will become tractable once Meyniel et al.’s account is modified to account for strongly embodied noetic feelings of confidence. Before either can be argued for, however, it is important to notice how Meyniel et al.’s theory of confidence neglects this embodied dimension to confidence.

From our discussion above, it is clear that subjective confidence is a cerebrally-produced signal, the upshot of cognitive control mechanisms, one which does not depend upon any extracerebral process for its production. As such, Meyniel et al. offer, what has been described as, a traditional cognition view about the nature of conscious confidence, which, as discussed earlier (Chapter 2), encounters issues when accounting for recent empirical evidence about how noetic feelings are conscious expression of emotion. So, what will be proposed is an augmentation to Meyniel et al. theory to account for a fifth explanatorily distinct kind of confidence, namely a strongly embodied noetic feeling of confidence, which, as it turns out, offers a plausible path forward to solving the problem of the origin of self-knowledge.
3. Strongly Embodied Noetic Feelings of Confidence and the Theory of Predictive Processing

The goal of this section is to demonstrate how Meyniel et al.’s account of metacognition, as it has been interpreted above (Chapter 5 and Section 2), can be modified to describe embodied metacognition, in which strongly embodied noetic feelings of confidence play a pivotal role in enabling and facilitating control over cognition and behavior. In so doing, Meyniel et al.’s account will merge with a prominent view from interoception-based research in emotion about the role of the limbic system in monitoring and controlling cortical areas (see Chanes and Barret 2016). In the final section, this hybrid view, which essentially describes the neurological underpinnings of embodied metacognition, will be argued to provide us with crucial tools for making tractable the ultimate problem on the origin of self-knowledge.

In discussing their theory of affect and emotion, Chanes and Barret (2016) provide us with a framework for understanding the role of the limbic system in producing conscious metacognitive feedback signals. It starts by understanding how implementing predictive coding principles within the structural model of corticocortical connections reveals that the direction of predictions and errors is determined by the laminar structure of the relevant areas: “predictions flow from less to more laminated cortices and prediction errors flow in the opposite direction” (ibid: 2; see Barret and Simmons 2015). Since “cortical limbic areas have the simplest laminar structure in the neocortex”, Chanes and Barret hypothesize that the limbic system is “at the top of the predictive hierarchy in all cortical systems” (Chanes and Barret 2016: ibid).

This hypothesis is based on the Barbas’ structural model of corticocortical connections (Barbas 1997, 2015). Appealing to the principle of systematic structural variation, that changes in laminar structure are systematic and not random, this model accounts for the strength and topography of connections between cortices and subcortical structures. The function of connections between structures can be determined by examining pathway projections and thereby the role of these structures in the overarching cognitive economy can be determined as well. Connections are thus categorized as belonging to either one of three kinds. Feedforward connections are narrow and project to areas away from the sensory periphery, while feedback connections are the opposite, dense and project to areas close to the sensory periphery. Finally, lateral connections link areas with small differences in structure and tend to be more distributed than the other two.
Drawing upon the neuroanatomy of rhesus monkeys (see Barbas 1986, 1993), Barbas’ model describes feedforward connections as originating from areas with more elaborate laminar structure than the destination, while the opposite relationship describes feedback connections (i.e. having less elaborate laminar structure than the destination). In particular, limbic areas were discovered as receiving feedforward projections from more elaborately laminated areas regardless of their location within the cortex (Barbas 1986). Applying the lessons of these studies to the Bayesian Brain Hypothesis and the predictive processing framework yields Chanes and Barret’s proposal: the limbic system sits atop the cognitive hierarchy, issuing predictions about causes of afferent information to all cortical systems through feedback connections, while also receiving prediction errors from those areas through their feedforward connections. Thus, limbic areas shape and constrain how this information influences executive control over cognition and behavior.

To relate this back to our discussion of Meyniel et al.’s account of subject confidence above, the cortices of the limbic system thus ought to be considered critical higher-order cortices that monitor and evaluate processing in lower-order cortices, crucial for the production of conscious metacognitive feedback signals. Due to its anatomic location, its position within the hierarchy, and its strong interconnectivity, “the limbic cortices create a highly connected, dynamic functional ensemble for information integration and accessibility in the brain.” Thus, Chanes and Barrett hypothesize further, the limbic system contributes to creating a “unified conscious experience” (ibidem). The limbic system is “in a privileged position to contribute to the neural basis of conscious access [that] may provide a “workspace” for conscious experience” (ibid: 6). In sum, the so-called ‘limbic workspace’ is the embodied, affect-laden and emotional seat of consciousness positioned at the top of the predictive hierarchy determining priority, relevance, and meaning at the personal (or animal) level of description.

With its role in regulating the autonomic nervous system, the sympathetic and parasympathetic nervous systems, as well as its connections to the prefrontal cortex, the limbic system might well be regarded as the central hub of embodied metacognition. If this is correct, and readout mechanisms are integrated into the limbic workspace, taking distributional confidence as input and yielding summary confidence as output, it places the scalar representation of summary confidence in a prime position to be “translated” into the language of affective bodily arousal (see Chapter 2). Thus, the limbic system, by processing bodily afferents, will shape and constrain the consequences distributional confidence will exert on behavior. Plausibly then, the limbic system transforms subjective confidence into strongly embodied noetic feelings of confidence.

Recall in previous chapters (Chapters 2 and 4) the recent empirical evidence that exposes an intricate relationship between noetic feelings and bodily afferents, in particular those afferents
produced by cardio-respiratory activity. On the basis of these interoception-based studies (Chapter 2), as well as how extracerebral processing is described by the neurovisceral integration model (NVM) as enabling and facilitating conscious control over cognition and behavior (Chapter 4), it is plausible that, given its role as the higher-order network, the limbic system transforms subjective confidence into strongly embodied noetic feelings of confidence through the interoception of homeostatic changes, in particular those changes tied to the cardio-respiratory system indexed by heartrate variability (HRV).

Conscious noetic feelings of confidence will thus exhibit robust informational correlations with evaluations of dynamic activity in lower-order cortices, so that the mental states characterized by the bodily afferents that constitute noetic feelings of confidence will exhibit valence and degrees of arousal described by scalar representations produced by this process of evaluation. Of course, such scalar representations are summary confidence (either (ii) or (iii) described by Meyniel et al.), which are nothing less than precision estimations of (i) distributional confidence, which is, in turn, the product of dynamic activity in the lower-order cortices. As such, a rough sketch of how distributional confidence transforms into strongly embodied noetic feelings of confidence is provided, one which transforms Meyniel et al.’s account into an embodied account.

As a result, conscious noetic feelings of confidence motivate and guide executive control over cognition and behavior, enabling the subject to improve cognitive performance to better correspond to the evaluations of the metacognitive control structure in the limbic system. Of course, this is a cyclical process, since the motivated acts of executive control will ultimately produce their own layers of feedback, shaping further input to the limbic system to be evaluated and readout as noetic feelings. Plausibly then, the limbic system is the central hub for transforming distributional confidence into strongly embodied noetic feelings of confidence.

As of now, it has been demonstrated how Meyniel et al.’s account of evaluative metacognition, with its solely cerebral confidence signals, can be easily modified to describe embodied metacognition, which produces strongly embodied noetic feelings of confidence, comprising a fifth explanatorily distinct kind of confidence due to its dependence on the limbic system and extracerebral processes. Of course, the argument is not that Meyniel and colleagues must accept this exact cognitive architecture of the limbic workspace to account for the role of the extracerebral body in the production of conscious confidence and uncertainty. Rather, the claim is that an appeal to the limbic workspace complements their theory, since it enables it to account for the empirical evidence from interoception-based studies on metacognition discussed above (see Chapters 2 and 4). So, with this modified account now on the table, let us discuss why it should be preferred over the unmodified, non-embodied account.
As already mentioned, Meyniel et al.’s account is conceptually similar to “traditional” accounts discussed above in Chapter 2, in that conscious confidence is explained by appealing solely to a central and cerebral cognitive control mechanism, such that no extracerebral process is appealed to in order account for how confidence is produced, nor how it becomes conscious, nor how it motivates and guides executive control over cognition and behavior. It was demonstrated above how such an account runs counter to evidence from studies that show the extracerebral body playing an enabling and facilitating role in this process of cognitive control.

Recall from Chapter 2 that several recent interoception-based studies in metacognition have revealed the degree to which the extracerebral body is implicated in this process of executive control. In particular, it was argued that a Neo-Jamesian theory ought to be adopted with respect to noetic feelings, including noetic feelings of confidence, such that, what we feel, when we feel confident, is the product of the interoception of bodily afferents, potentially those produced by the cardio-respiratory system.

Of course, a proponent of the traditional view might postulate two kinds of confidence, a non-embodied and an embodied kind. But this line of reasoning was the precise target of criticism in Chapter 2. There it was discussed how this proposal stands in need of empirical support and it was argued that several recent studies from interoception-research in metamemory demonstrate the essential role that the body plays in the cognitive control process through its production of bodily afferents that constitute noetic feelings, including feelings of confidence, which, in turn, motivate and guide executive control over behavior. Thus the empirical evidence warns against adopting a view of two kinds of confidence (an embodied and a non-embodied kind).

One study is particularly pertinent given our discussion of the theory of predictive processing. Recall how Allen et al. (2016) observed how the production of unanticipated bodily afferents can alter (even invert) computational biases in cognition and thus influence subjective reports of confidence. Allen et al. (2016) theorize that the extracerebral body plays a foundational role in shaping conscious control by influencing neurocomputational priority settings, that is, influencing how precision estimations of distributional confidence impact behavior.

This evidence points in favor of an embodied account of Meyniel et al.’s theory of confidence and thus supports the claim that the limbic system subserves executive control over cognition and behavior through its interoceptive processing of bodily afferents. As such, an embodied metacognitive mechanism might well have the proverbial ‘final say’ in how relevant patterns of distributional confidence come to influence executive functioning.
Moreover, recall from Chapter 4 how NVM describes the extracerebral body as playing an enabling and facilitating role in executive control, a multifaceted process, crucial aspects of which can be observed by measuring both task-related and resting-state HRV, cardiovascular changes that, studies show (see Fiacconi et al. 2017; Meessen et al. 2018, respectively), correlate in intriguing ways with subjective reports of noetic feelings. According to NVM, increases in task-related HRV, as observed during episodes characterized by noetic feelings relative to episodes not so characterized (see Fiacconi et al. 2017), are theorized to index the activation of crucial neurovisceral control mechanisms that enable and facilitate executive control over cognition and behavior.

Meanwhile, also discussed in Chapter 4, greater resting-state HRV – as observed in subjects who perform better in confidence-based tests of metamnemonic awareness relative to subjects with lower resting-state HRV (Meessen et al. 2018) – is described by NVM to index enhanced inhibitory capacities, the deployment of which is theorized to be enabled by the activation of crucial neurovisceral control mechanisms that, in turn, enable and facilitate executive control. Thus, this evidence supports an embodied construal of Meyniel et al.’s theory of confidence, in which extracerebral processes are crucially implicated in the production of conscious confidence.

Further support for an embodied account can be obtained if it can be shown how an appeal to the production of confidence signals in this manner will provides us with crucial tools for making tractable the problem of the origin of self-knowledge. Though this will be the topic of the next section, let us consider the initial motivation for this proposal.

By guiding distributional confidence through neural circuitry tied to the regulation of homeostatic variables (a process known as ‘allostasis’; see Stirling 2012), by coupling neurocomputational activity to cardio-respiratory activity, nature has ensured distributional confidence, approximating Bayesian confidence, will become integrated into the subject’s embodied self, thus comprising cornerstones of her foundation for selfhood (see Seth 2013, 2015). So, the guiding thought is that by becoming embodied, confidence signals manifest in a highly auspicious manner, pivotal to explaining how subjects acquire self-knowledge.

In particular, strongly embodied noetic feelings will demand subjects not only explore these feelings, should they be pleasant, and the epistemic opportunities revealed by them, but also, should they be unpleasant, come to resolve these feelings in a manner that is emotionally satisfying. The end result of undergoing confidence, and its inverse, uncertainty, in an embodied format are thus robust information-seeking behaviors that plausibly comprise the initial foundation for practices that constitute a collective epistemic agency. In this respect, evidence for such robust psychological states in non-human animals was already introduced above (see Chapter 5, Section 3).
At this point, I believe some speculation is warranted, so long as, that is, such speculation also envisions empirical predictions that can be plausibly tested, thus generating hypotheses for future research, a topic that will be addressed at the end of the present work (see Conclusions). Plausibly, strongly embodied noetic feelings of confidence are poised to manifest externally, transforming into ostentatious behavioral expressions that can be readily detected, monitored, and regulated by other individuals in a subject’s socio-cultural environment. Subjects endowed with such cognitive systems (i.e. higher order cortices that monitor and evaluate the processing of other cortical areas, which are themselves tied to both prefrontal levels of control as well as the peripheral and autonomic nervous systems) are therefore in a prime position for regulating behavior and the behavior of others in a manner sensitive to what these expressions ultimately represent, namely precision estimations of patterns of distributional confidence (see Section 4).

Together these two features comprise two regulatory practices, one intra-individual and one inter-individual, that could plausibly transform the socio-cultural environment into one in which individuals are treated as more or less confident in their abilities, as believing more or less in themselves (for a discussion of this intra-individual regulatory practice as an emotional maturation process, see Chapter 4). In other words, these subject will then become treated as holders of self-beliefs and as proto-epistemic agents, thus shaped by nature to discover themselves as self-believers. Therefore, distinguishing strongly embodied noetic feelings of confidence from mere conscious confidence signals provides us with crucial tools for building a bridge between the phylogenetic and the ontogenetic origin of self-knowledge, thus making significant inroads into solving the overarching problem of the origin of self-knowledge. The challenge is now to make a plausible case for how noetic feelings of confidence can ultimately account for capacities for self-knowledge in a manner that mere conscious confidence may struggle to do.
Chapter 6: How Feeling Our Hearts Enables Us to Feel Our Minds


This section will seek to provide a plausible framework for making inroads into the problem of the origin of self-knowledge by bridging the phylogenetic origin, as described by theories of metacognition, to the ontogenetic origin, as described by the social-scaffolding view about how mindreading capacities emerge during development through the process of enculturation. The proposal, already introduced above, involves appealing to two crucial features of embodied metacognition. Though the goal of this section is to argue for how these features plausibly pave the way for solving the problem of the origin of self-knowledge, let us first review the exact details surrounding these two features.

The first feature of embodied metacognition is its propensity to produce mental states characterized by strongly embodied noetic feelings. These mental states are robust psychological states characterized by conscious emotion. The argumentative crux is thus that these emotional states are well suited to promote rich information-seeking behaviors due to their status as emotional states. This is one half of what appealing to confidence as a feeling can easily account for that an appeal to non-feeling-based conscious confidence struggles with.

The other half are external manifestations of confidence, which an appeal to strongly embodied noetic feelings can easily account for due to their nature as feelings, while an appeal to non-feeling-based confidence signals struggles to explain. Specifically, embodied metacognition produces external manifestations of confidence, behavioral expressions that can be detected, monitored, and regulated by other individuals in the subject’s socio-cultural environment, a crucial tool for building a bridge between the phylogenetic and the ontogenetic origin of self-knowledge.

Together these two features lay the foundation for those practices that could plausibly go on to form a collective epistemic agency. To refer to a previous simplification of the process (Introduction), strongly embodied noetic feelings are equivalent to the chicken-like creature that produced the egg (the socio-cultural practices and collective epistemic agency) that produced the chicken (self-knowledge). Of course, this is under the assumption that the social-scaffolding view about the ontogenetic origin of self-knowledge is on the right track.

Let us begin by building the case for how strongly embodied noetic feeling of confidence and uncertainty are well suited to promote rich information-seeking behavior. Consider how good it feels to arrive at a solution, how good it feels to be confident in decision-making. Consider also
how irritating confusion can be, the cognitive dissonance created by nonsense. Pleasant feelings drive us not only to act on the affordances presented, but also to explore them, to put ourselves in similar situations in order to feel them again. Meanwhile, unpleasant feelings motivate us to resolve this unpleasant sensation in an emotionally satisfying manner, that is, to replace the unpleasant feeling with a pleasant one that signals the former’s satisfactory completion.

As argued for above (Chapter 4), due to the demands of an unforgiving environment and how these mental states represent proto states of self-belief, such states give rise to an emotional maturation process that approximates the updating of beliefs in light of evidence and counterevidence. Thus, this process of exploration and resolution emerges as the result of noetic feelings specifically as strongly embodied conscious expressions of emotion.

Thus, when feelings represent cognitive affordances, as is the case with strongly embodied noetic feelings, robust behaviors emerge around exploring and resolving epistemic opportunities and epistemic threats. These behaviors ought to be construed as forms of ‘sense-making’ due to how the subject brings forth the informational significance of such environmental (or internal) features (see Thompson and Stapleton 2009). Such behaviors are referred to as ‘epistemic foraging behaviors’, theorized to underlie much of what ought to be considered intelligent behavior (see Friston et al. 2015). These behaviors explore and exploit safe forms of uncertainty, e.g. possible locations of a hidden food source, which are theorized to provide animals with crucial survival advantages because of how these behaviors will ultimately develop skills around mitigating and coping with dangerous forms of uncertainty, e.g. food scarcity.

Though it is unclear whether these behaviors are fundamental to all living creatures or whether only some manage to engage them, what is clear, is that doing so will provide those subjects with survival advantages as they become better able to minimize prediction error (free energy and entropy) in the long run. This is not to say that epistemic foraging requires strongly embodied noetic feelings. Rather, it is to say that strongly embodied noetic feelings are particularly well suited, as strongly embodied, to produce robust epistemic foraging behaviors.

Consider again the fascinating results from the uncertainty monitoring paradigm, specifically how the metacognitive performance of humans, dolphins, and primates differs from that of ideal observers, in particular how these animals (ourselves included) will choose to opt-in where they ought to have opted-out, that is, if the goal was to maximize immediate rewards. Thus, their decision-making behavior does not aim to maximize immediate rewards. Rather, it aims to explore this safe kind of uncertainty and, in so doing, determine the exact threshold at which point successful visual discrimination is no longer tenable.
This is indeed sophisticated behavior. Plausibly, the subject is teaching herself about her own limitations, pushing herself to learn more about her own cognitive system, whether she knows she is doing so, or not. Crucially, as was already introduced (Chapter 4), comparative psychologists suggest this somewhat counterproductive behavior is explained by the subject’s need to resolve the uncertainty produced by engaging in the task in an emotionally satisfying manner (see Smith et al. 1995). It is plausible that these rich epistemic foraging behaviors, exposed by the uncertainty monitoring paradigm, are motivated and guided by robust emotional states. What is more, if the present account is correct, these mental states, even those of non-human animals, are characterized by strongly embodied noetic feelings, a speculative point for sure, but one that will be revisited at the end of the present work and demonstrated to yield testable predictions (see Conclusions).

Recall from above (Chapter 4) how strongly embodied noetic feelings were argued to characterize mental states that ought to be construed as precursors to full-blooded states of self-belief. This was because of how these mental states represent (in the de re mode) neurocomputational changes that describe the subject’s own capacities for action, as well as how these states aim at truth, while also possessing mind-to-world direction of fit, typical of belief.

Also recall how the ultimate result of subjects coming to occupy these mental states in their engagement with the world, transitioning from states of high to low confidence and vice versa, was argued to reflect an emotional maturation process that approximates the process of updating beliefs in light of evidence and counterevidence. As such, an appeal to the extracerebral body and its role in producing emotional experiences in the form of strongly embodied noetic feelings of confidence is crucial to account for this intraindividual regulatory process.

Albeit speculative, such intraindividual regulation centered around confidence and uncertainty with respect to one’s own capacities can potentially serve as the initial foundation for those key practices that structure a collective epistemic agency. Intuitively, demands issued by members of the socio-cultural environment about individuals regulating their own behavior (however implicit these demands are) can only be met if those individuals are first in a position to monitor and control their own behavior.

This is not to say the capacity for intraindividual regulation must precede or serve as a precondition for the capacity to issue such demands (in reality, the origin story of such capacities might be far more convoluted). But it is to say that the capacity to meet these demands presupposes basic capacities for intra-individual regulation, capacities which could potentially be enabled by mechanisms for embodied metacognition. Of course, this is no small feat, as the capacity to meet
such demands is an essential part of any robust species of collective epistemic agency, the origin of which is included in our ultimate explanandum.

Let us return now to Meyniel and colleagues’ theory of confidence and consider its unaugmented version in terms of how it would account for metacognition. Recall that the unaugmented account appeals to solely-cerebral subjective confidence signals. What is missing is thus the link to rich epistemic foraging behaviors that are motivated and guided by robust psychological states characterized by emotional arousal. Perhaps, Meyniel et al. can connect mental states characterized by subjective confidence to epistemic foraging behaviors, but it is unclear how it exactly this might be achieved without appealing to emotional valence, arousal, or the experiential character typical of emotional experience, while also describing a conscious process.

Of course, one proposal would be to appeal to the neuromodulators of reward systems to account for the motivation and guidance required of epistemic foraging behaviors (see Yu and Dayan 2005; Parr and Friston 2017). However, such systems are theorized to subserve basic cognitive mechanisms, such as those involved in reinforcement learning, and are not generally theorized as sufficient for executive functioning (Phillips et al. 2008; Schultz 2010). For example, a seminal study by Braver et al. 1999 found that dopaminergic systems can merely gate the access of representations into working memory, in effect modulating attention. Thus, in order to account for the right sort of functional role sufficient to explain such sophisticated behaviors as described above, Meyniel et al. would likely need to appeal to the output of reward systems as intertwined with the global workspace in line with their theory (see above).

The upshot would be solely-cerebral conscious signals that could potentially serve as analogues to emotional arousal, thus potentially serving to motivate and guide sophisticated epistemic foraging behaviors that play a crucial role in enabling intra-individual regulatory practices, crucial materials for building the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge. But notice that such signals readily lend themselves to being construed as anti-Jamesian sorts of feeling that recent empirical evidence suggests ought to be looked upon with a due amount of skepticism (see Chapter 2). Thus, the traditional view encounters challenges whenever appealing to non-embodied signals in explaining how such signals are produced and fulfill their role in the overarching cognitive process. As it turns out, the extracerebral body and emotional experience might be essential to explaining executive control.

Let us now build the case for how strongly embodied noetic feelings of confidence and uncertainty are well suited to manifest externally, become expressed in ostentatious behaviors that can be readily detected, monitored, and regulated by other individuals in the subject’s socio-cultural
environment. Recall Proust’s theory of the origin of epistemic normativity, specifically how she describes the correctness conditions of mental states characterized by noetic feelings as approximating accuracy conditions in virtue of their relationship to cognitive fluency (Chapter 1). Recall that this is because of how fluency, more often than not, approximates accuracy.

This simple but profound realization about the relationship between fluency and accuracy has another crucial side to it, one that is, likewise, as intuitive as it is non-trivial. Namely, confidence, when well placed, approximates competence. This means, behaviors motivated and guided by mental states characterized by strongly embodied noetic feelings of confidence will approximate competent behaviors, and, intuitively, any species that learns to detect behavioral expressions of confidence will gain a crucial evolutionary advantage because of this relationship.

Consider decisions about the assignment of various social roles, such as those that depend upon both collaborative and competitive interaction, like those of foraging, hunting, and mating (see Call and Tomasello 2008). Since confidence approximates competence, the external manifestations of confidence are prime candidates to be detected and monitored by subjects for the purpose of inter-individual regulation, that is, be readily used to determine an individual’s role within the socio-cultural environment, and thereby shape the social identity of its members.

Clearly then, subjects who depend upon others for survival will obtain survival advantages if some measure of competency can be drawn upon when making decisions such as these. Admittedly, appeals to adaptive fitness are only a prima facie reason to accept a proposal, but more specific predictions emerge, which are discussed at the end of the present work (Conclusions).

Take note that the origin story of self-knowledge told by the present work does not require the cognition of proto-epistemic subjects to possess joint-intentionality, the capacity to form intentions structured by the expectations of other individuals (Tomasello 2019). Rather, it is a story about a primitive form of epistemic accountability, in which the behavior of subjects is sensitive to the potential consequences of their own expressions of confidence, as well as, at times, sensitive to the expressions of confidence made by other subjects.

As discussed throughout the present work, subjects need not entertain mental states with sophisticated structure that would demand joint-intentionality in order for their behavior to be sensitive along these proto epistemic dimensions. Instead, subjects need only undergo noetic feelings with strong informational correlations to neurocomputational changes, on the one hand, and come to undergo these as conscious expressions of emotion and express them in the form of ostentatious behaviors, on the other.
Thus, external expressions of internal confidence, manifestations of strongly embodied noetic feelings, can plausibly bridge the phylogenetic origin of self-knowledge to its ontogenetic origin, that is, so long as the social-scaffolding view is on the right track. A society, however primitive, that adapts in such a way so as to regulate social-role assignments based on externally observable expressions of internal confidence becomes one that treats its members as proto epistemic subjects, thus treating them as dealing in proto beliefs about their own abilities.

Consider again the unaugmented version of Meyniel et al.’s account and ask whether it has the tools to build a similar bridge. Does subjective confidence have corresponding external manifestations? One proposal is that subjects who experience subjective confidence will likely execute their actions with a specific behavioral profile, characterized by speed, accuracy, and consistency. In other words, if done confidently, the act will be done faster, more accurately, and more consistently than those that are done with less confidence. As such, these features reflect external manifestations of subjective confidence, and so Meyniel et al.’s unmodified account is not completely without the tools for constructing a bridge between the two origin stories.

However, this approach is not without its disadvantages relative to the embodied approach. The first is that this behavior profile characterized by speed, accuracy, and consistency is tightly bound to the actual performance of the action in a manner that the embodied approach is not. This is because these three characteristics will tend only to manifest while the action is performed or while the social role is fulfilled. Meanwhile, embodied expressions of confidence such as the raising of eyebrows to signify confidence and the unfurling of a frown to signify uncertainty, can be detected before, during, as well as after the performance of an action or the fulfillment of a social role, thus making embodied expressions far more prominent and far more useful as a target for inter-individual regulation. Again, adaptive fitness may only be a prima facie reason, but there are empirical predictions that fall out of this proposal (see Conclusions).

Second, the behavioral profile characterized by speed, accuracy, and consistency is sorely lacking in nuance that the embodied approach is poised to accommodate. Though these three characteristics can be measured along some scale, and so do admit of degrees, there is much less specificity. Consider the range with which confidence and uncertainty can be expressed when manifest by all manner of ostentatious actions, from subtle expressions, such as raising of one’s eyebrows, to more vulgar displays of confidence, such as vigorously jumping in place.

These manifestations are so multifaceted that they might plausibly acquire contextual appropriateness to them, so that raising eyebrows is how one expresses confidence in some cases, e.g. during mating rituals, but jumping in place is how one ought to express confidence in other
cases (during foraging). Together these multifaceted expressions can yield more adaptive behaviors and form a robust collective epistemic agency, while it is unclear whether behavior profiles characterized by speed, accuracy, and consistency are able to achieve such an array of expressions.

In human subjects, one can easily read subjective confidence, and its inverse uncertainty, from a person’s bodily language. Consider the young school child who frantically raises her hand whenever the teacher asks the class a question, demonstrating just how confident she is that she knows the answer. Strongly embodied noetic feelings are thus a window into the inner life of other subjects as self-knowing agents (whether or not they know it or we know it as that).

By detecting another person’s levels of confidence, we detect what this person thinks about themselves, their own behaviors and beliefs, information that can be used to guide our thinking about them, all of which approximates capacities for mindreading. As such, strongly embodied noetic feelings reflect plausible precursors to those features that are detected whenever deploying mindreading capacities, possibly reflecting a phylogenetic origin to such capacities.

Of course, with the right sort of cognitive machinery it is plausible that cerebrally-generated signals can produce such nuanced bodily expressions without thereby requiring these signals be perceived via the interoception of bodily afferents. But, again, the empirical evidence discussed earlier suggests this is indeed what is going on in these cases, namely, metacognitive awareness and interoceptive awareness are supported by a common mechanism, and signals conveying confidence and uncertainty are made conscious through the interoception of bodily afferents (see Chapter 2).

Notice however that if this is the response – namely that an appeal is made to centrally-mediated feedback signals as causally responsible for behaviors of confidence and uncertainty crucial for building the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge – then the approach begins to assume the spirit of an embodied approach to metacognition. Since, even if these bodily expressions are produced as the result of non-embodied central signals, these bodily expressions are thus cognitive components in their own right, used as input for further cognitive processing either by the subject herself or by other individuals in the socio-cultural environment.

Obviously, this is distinct from an appeal to bodily afferents, but appealing to bodily expressions as crucial inputs to the interindividual regulatory process amounts to appealing to extracerebral processes as something more than mere peripheral components. Such an appeal
would thus make Meyniel and colleagues’ case a variant of an embodiment approach. It is just a very weak embodied approach, which does not involve an appeal to interoception but behavior.\footnote{Notice, however, that it is not strictly a behaviorist view as this construal of Meyniel and colleagues’ view would continue to appeal to internal states of confidence and uncertainty.}

Thus, regardless of whether there is no agreement on the exact configuration of their enabling mechanism (i.e. on whether or not this mechanism includes interoception and bodily afferents), there might be nonetheless common ground around the proposal that behavioral expressions of confidence and uncertainty comprise crucial ingredients in bridging the two theories about the origin of self-knowledge. But behavioral expressions of confidence and uncertainty have not been placed at the forefront of the research, so there is little empirical evidence which tells us whether this proposal is on the right track.

Studies by Bahrami and colleagues (2010) and (2012) tested the degree to which confidence shapes social interaction under uncertainty, but these studies were based on verbal interaction and not bodily expressions of confidence. Meanwhile, Mahmoodi et al. (2013) controlled for non-verbal communication, but this was measured by having subjects share their confidence levels with each other through an onscreen interface, not by observing one another’s behavior.

So before concluding this section, let us entertain an example that might reinforce the common-ground proposal that behavioral expressions of confidence and uncertainty might comprise key materials in the bridge between the origin stories of self-knowledge. Consider again the dolphin Natua who performed so well in the uncertainty monitoring paradigm that he displayed sophisticated epistemic foraging behaviors around exploring the thresholds of ambiguity.

As Smith and colleagues (1995) noted, Natua’s lack of confidence during these borderline cases was manifested externally. For example, he would open his mouth wide and shake his head from side to side. Such expressions are ostentatious and so readily lend themselves to being detected and monitored by other individuals, just as we should predict if the current proposal is correct. These behaviors had to come from somewhere and so likely emerged in the service of some adaptation. It is possible then that such behavioral manifestations comprise essential materials to the foundation upon which inter-individual regulatory practices of an epistemic kind are built, and, with it, an origin for capacities for acquiring self-knowledge. But the question remains: are these behavioral expressions the products of centrally-produced signals or the consequence of strongly embodied feelings?

Of course, the present work has argued it is safe to say that, in our case, mental states characterizing such episodes are infused with strongly embodied noetic feelings. Recall, of course,
the empirical evidence of the crucial role that bodily afferents play in moderating and potentially mediating confidence and uncertainty (see Chapters 2 and 4). But in the case of non-human animals we do not have any evidence of this because such claims have not been tested yet. That is, as far as I know, it has not been asked yet whether non-human animals even have interoceptive capacities, and so it has not been investigated whether metacognitive awareness shares a common mechanism with interoceptive awareness in non-human animals. Thus, a great deal about this story will need to be completed before such conclusions can be drawn (see Conclusions).

To conclude, consider how, even for us humans, the path of self-discovery paved by strongly embodied noetic feelings is far from complete. Consider how we now teach judges about how their sentencing can become stricter when feelings of hunger become confused with feelings of incredulity or impropriety (see Corbyn 2011). Consider how sport coaches have started teaching young athletes about how the nervous arousal felt before a big game does not actually mean a lack of confidence but rather the opposite, since this arousal is actually the feeling of being excited and ‘in the zone’ that ought to signal confidence in the ability to compete (see Heyes et al. 2020).

Consider also how we are only recently started teaching people, adults included, not to follow their gut feelings when making decisions, that noetic feelings often result from implicit decisional biases that can actually hinder good decision making. What is potentially more, consider the pandemic of misinformation poisoning our current age, posing a threat to rational discourse and democracy, how the influence of such non-information, masquerading as information, might be explained in part by appealing to its propensity to exploit the influence that noetic feelings have on our decision-making behavior. In many ways, humans are still on the path of discovery about how noetic feeling shape our capacities for self-knowledge, and, in many respects, our discussion about the nature of noetic feelings in this present work is one of this path’s many steppingstones.

5. Conclusion

If the above analysis is correct, it means it is deeply plausible that the mental states animals are in when choosing to opt out during uncertainty monitoring tasks are states that ought to be characterized as embodied experiences, embodied summaries of distributional confidence, calculated and emerging out of the depths of the animal’s cognitive machinery, states that ensure the animal feels its self as confident. We are thus warranted in describing these mental episodes, these embodied noetic feelings of confidence and uncertainty, as reflecting precursors to full-
blooded states of self-knowledge and so warranted in describing this process as a possible origin for self-knowledge.

As we move forward with the strategy of applying the theory of predictive processing to the domain of metacognition, we would benefit from keeping in focus the latter’s original explanandum: how to make sense of the fact that some animals demonstrate advanced decision-making skills that involve the mitigation of uncertainty and the utilization of confidence for the benefit of improved discriminatory performance on par with that of adult humans. The guiding intuition has been, and I believe should continue to be, that these animals are able to make these decisions not because of advanced analytic or conceptual capacities, but rather, much as we do, as the result of undergoing a specific class of affect, noetic feelings, signatures of extracerebral activity that has, through the course of natural evolution, come to make salient the patterns of distributional confidence occurring within and across cognitive operations.
In writing this dissertation, I examined the case for evaluative metacognition as a possible phylogenetic precursor to more sophisticated forms of thinking about thinking, such as those made possible by deploying capacities for introspection, mindreading, and metarepresentation, and thus a possible phylogenetic precursor to self-knowledge. In particular, I examined the role of noetic feelings in evaluative metacognition, such as the feeling of confidence, certainty, and the feeling of knowing, as it pertains to executive functioning and provided an empirically-informed account of the underlying mechanism that explains how noetic feelings motivate and guide epistemic acts, like the act of seeking out justification for belief, judging some content to be accurate, or perceptually discriminating the larger of two visual stimuli.

I concluded that, to the extent that they are discussed in the literature, noetic feelings come in various degrees of extracerebral involvement, from absolutely zero bodily involvement to quite a significant amount (see Chapters 2 and 4). In terms of understanding the origins of self-knowledge, however, I argued that noetic feelings ought to be appealed to only to the degree that they are strongly embodied. As strongly embodied, noetic feelings are intricately intertwined with bodily activity, specifically that activity that is indexed by changes in heart rate variability, bodily activity which, in turn, is produced by and plausibly represents (in the de re mode of reference) changes in neurocomputational activity, specifically those changes produced by cognitive control structures. Accordingly, what you feel, whenever you feel a noetic feeling, are bodily changes with strong correlations to changes in neurocomputational activity, so that we might say that, to the degree that noetic feelings are precursors to self-knowledge, *we feel our hearts to know our minds*.

One prime example of noetic feelings is the feeling of confidence that motivates an agent to judge that her belief is true and justified. I assumed that noetic feelings, like the feeling of confidence, reflect, at best, steppingstones on the phylogenetic path to advanced forms of self-knowledge. Indeed, in order for self-knowledge to emerge as such, creatures must be embedded in sophisticated socio-cultural environments, in which practices have developed for treating individuals as bearers of belief states.

Crucially though, noetic feeling could very well play a pivotal role in explaining how such practices emerge. This is because of how noetic feelings, as strongly embodied, lay the foundation for both intra-individual and inter-individual regulatory practices that approximate the updating of
beliefs in the light of evidence and counterevidence. With respect to intra-individual practices, these were argued to reflect an emotional maturation process. With respect to inter-individual regulatory practices, it was argued that strongly embodied noetic feeling are poised to produce ostensible bodily expressions that can be readily detected and observed externally by others. The reason this aspect is so important to the role of noetic feelings in the origin of self-knowledge is because of how these bodily expressions can serve as bridges from subjects to other subjects, cluing others into the thinking of individuals, serving a rudimentary form of mindreading. This enables noetic feelings to serve as socio-cultural commodities to be negotiated with and thus shape the socio-cultural identity of individuals as proto epistemic agents.

That said, the present work did not provide a knockdown argument for rejecting the traditional approach to shoring up metacognition against the problem of triviality, other than by pointing out how it struggles to account for recent empirical evidence from interoception-based studies in metacognition (see Chapters 5 and 6). To assuage any remaining skepticism about the Neo-Jamesian theory of noetic feelings, more empirical research will need to be conducted. For example, improved techniques for measuring interoceptive sensitivity that assess interoceptive awareness via the detection of respiratory activity could be applied to study whether metacognitive awareness and interoceptive awareness share a common mechanism (e.g. Allen et al. 2021).

With respect to making inroads into solving the problem of the origin of self-knowledge, though the present work was likewise unable to provide a knockdown argument for rejecting the traditional approach, it was argued that an embodied approach has more tools at its disposal for solving this problem (see Chapter 6). In this context, appealing to the body is appealing to emotion, and this provides a robust framework for making sense of the complex psychological states that motivate and guide robust epistemic foraging behavior. Additionally, appealing to the interoception of bodily afferents provides a robust framework for accounting for highly-nuanced external manifestations of confidence and uncertainty, which could serve as crucial materials for building the bridge between the phylogenetic and ontogenetic origin stories about self-knowledge.

At this crucial junction, several more questions emerge that could be explored by future research. One central question concerns the origin of such sophisticated behavior as treating individuals as bearers of belief states. This is generally theorized to emerge from the need of a species to cooperate and compete within and across various social roles over limited resources. The question I raised here was whether some forms of cooperation and competition are better suited than others for the emergence of belief states as such. What was proposed above is that those forms of cooperation and competition that are sensitive to bodily expressions of confidence and uncertainty are better suited. This is due to how confidence and uncertainty, as mental states,
have various properties in common with robust belief states that adult humans entertain. This proposal was described above as potentially reflecting common ground between traditional (or weak embodied) accounts and strong embodied accounts (Chapter 6).

If this is correct, it raises questions about the phylogenetic origin and ontogenetic origin of such behaviors for expressing confidence and uncertainty. Have non-human animals developed behaviors for the expression of confidence and uncertainty? Do animals track these expressions in conspecifics and have they learned to respond to them adaptively so as to exploit them e.g. in the practice of assigning social roles? Can we observe an overlap between those animals that display advanced forms of self-awareness and such behaviors, e.g. in those animals that pass the mirror tests, succeed as uncertainty monitoring tasks, or engage in information-seeking behavior?

Of course, humans have developed various and often complex behaviors around expressing uncertainty and confidence. Consider the young child at school who vigorously raises her hand whenever the teacher asks the class a question, clearly demonstrating her confidence in knowing the answer. When do these behaviors begin to develop and how? Are some of these behaviors innate, perhaps as a consequence of our phylogenetic heritage? Do human children learn to track these behaviors in others? When do they start using these behaviors to assign social roles? All of these questions ought to be addressed by future research to determine the extent to which the proposal made by the present work holds.

Understanding when and under what conditions behaviors that express confidence and uncertainty have emerged both phylogenetically and ontogenetically will help us understand how creatures become treated as epistemic agents, as holders of belief states, as potential knowers. States of confidence and certainty are more basic than belief states since they do not require animals possess a concept of belief, but only require they have a means of expressing confidence and uncertainty. But the mental states of uncertainty and confidence are functionally analogous to beliefs, so they reflect prime candidates for proto belief states. Thus, understanding the origin of the behavioral expressions of confidence and uncertainty might help us to explain the origin of self-knowledge in creatures like us.


References


Comstock, Gary. 2019. What do we need to know to know that animals are conscious of what they know? *Animal Behavior and Cognition* 6 (4).


References


References


References


References


References


References


