5-1-2024

The Philosopher and the Beetle: Attention, Self-Representation, and the Fate of Predictive Processing

Frank Wotton

The Philosopher and the Beetle: Attention, Self-Representation, and the Fate of Predictive Processing

by

Frank H. Wotton IV

Under the Direction of Neil Van Leeuwen, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2024

ABSTRACT

I draw on work by John Perry to argue that the standard predictive processing theory of attention (SPPA) has the surprising consequence that any agent endowed with a predictive architecture and the capacity for attention can explicitly represent itself *qua* agent. I argue that this consequence presents us with a hard choice among four competing theoretical options: (1) preserve our intuitions about the self-representational capacities of non-human organisms and reject SPPA, (2) abandon our intuitions and preserve SPPA, (3) preserve our intuitions and preserve SPPA, but reject Perry's theory of self-knowledge, or (4) preserve our intuitions, preserve SPPA, and preserve Perry's theory, but qualify predictive processing such that *not all* adaptive self-organizing systems are PP agents.

INDEX WORDS: Attention, Self-representation, Predictive processing, Representation, Action

The Philosopher and the Beetle: Attention, Self-Representation, and the Fate of Predictive Processing

by

Frank H. Wotton IV

Committee Chair: Neil Van Leeuwen, PhD

Committee: Juan Piñeros Glasscock

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2024

DEDICATION

This thesis is dedicated to my parents whose love and support made all of this possible.

ACKNOWLEDGEMENTS

I first and foremost extend thanks to my thesis advisor, Neil Van Leeuwen, for his expert guidance throughout this project. I would not be the philosopher that I’ve become were it not for his sustained support and thoughtful feedback. The title of this thesis was also Neil’s idea. Some of the most significant developments of the thesis were also due to Juan Piñeros Glasscock’s famously thorough feedback. As a committee member, Juan went above and beyond in his mentorship, and for that I am sincerely grateful. Finally, I’d like to thank my undergraduate advisor, Richard Lauer, who was kind enough to take the time to read and provide feedback on the work of a student from days of yore.

TABLE OF CONTENTS

1 INTRODUCTION

Predictive processing (PP) claims that adaptive self-organizing systems aim to minimize discrepancies (called prediction errors) between an internal model of the causes of their sensations and their actual sensory input.[1] Prediction errors generate signals (called prediction error signals) that propagate up the perceptual hierarchy where they either cause the agent to update its model or generate an action that selectively sculpts and samples sensory data that confirm its model.

Under uncertain conditions, PP agents need a way to discern those discrepancies that are indicative of causal regularities in the world (what I shall from here on call "world-informative" discrepancies or prediction errors) from those that are not. According to PP, this role is fulfilled by precision optimization, a process that selectively increases the gain on statistically precise (reliable) prediction error signals and attenuates imprecise (unreliable) ones. The standard PP theory of attention (SPPA) identifies precision optimization with attention.[2]

I will argue that SPPA has the following consequence.

> **Self-awareness is Not Rare (SNR)**: If a PP agent (an agent endowed with a PP architecture) has the capacity for attention (according to SPPA), then that agent can explicitly represent itself *qua* agent.[3]

[1] See Friston et al. (2006), Friston & Stephan (2007), Friston (2009, 2010), Hohwy (2013), Clark (2013, 2016), and Parr, Pezzulo, & Friston (2022).

[2] See Friston (2009, 2010) and Feldman & Friston (2010).

[3] I use "*qua* agent" as an agent-neutral rendering of Castañada's "*qua* he himself." I say more about what this means in Sect. 2.

By “rare,” I mean that the proportion of agents who bear the capacity to represent themselves in this kind of way is small (e.g., it just includes fully-functioning adult human beings). (SNR) is surprising because most of us have the intuition that most non-human organisms cannot represent themselves, or if they do, they do so only in a very weak sense (e.g., accidentally or implicitly). Intuitively, the self-representational capacities of most non-human organisms end here. Dennett expresses this sentiment when he says,

> The hermit crab is designed in such a way as to see to it that it acquires a shell. Its organization, we might say, *implies* a shell, but the crab does not in any stronger sense *represent itself* as having a shell. It doesn’t go in for self-representation (Dennett, 1991, p. 417).

We might summarize this intuition as the intuition that self-awareness *is* rare. I think many of us have this intuition. It also straightforwardly contradicts (SNR), and that is why I think (SNR) is surprising. If SPPA is correct, then the philosopher, the beetle, and everything in between do in fact “go in for self-representation”—at least to a greater extent than we might have thought. In particular, they “go in” for explicit self-representation. And given that explicit self-representation is a building block for the more sophisticated forms of self-awareness in which philosophers and scientists alike have historically been most interested, this is theoretically important. It means that everything from lowly beetles all the way up to philosophers are a step closer to those more sophisticated forms of self-awareness than we might have thought. With that said, my aim in this paper is solely to show that SPPA has this consequence. Aside from a few brief suggestions in Sect. 5, I leave it to future research to decide what to do about it.

My basic argument is as follows.

> (P1) For any PP agent *S*, if *S* has the capacity for attention (according to SPPA), then *S* has a fallible self-model.

> (P2) For any PP agent *S*, if *S* has the capacity for attention, then if *S* has a fallible self-model, then *S* can explicitly represent itself *qua* agent.

> (SNR) For any PP agent *S*, if *S* has the capacity for attention, then *S* can explicitly represent itself *qua* agent.

(P1) relies on the assumption that all PP agents have self-models. Drawing on work from John Perry, in Sect. 3.1 I argue that all PP agents, regardless of whether they have the capacity for attention, have self-notions and that having a self-notion suffices for having a self-model. The self-model inherits its fallibility from the fallibility of the self-notion (since having a self-notion is a way of having a self-model).

(P2) is not obviously true, but it follows from SPPA. An agent that has a fallible self-model is vulnerable to a special class of prediction error that I call *self-prediction error*. These prediction errors are discrepancies involving self-representations in the perceptual hierarchy, and agents that have the capacity for attention must optimize the precisions of those prediction errors in a process that I call *self-attention*. Crucially, self-attention requires that the agent represent itself in attention *qua* agent, which means that agents that have the capacity for attention can represent themselves in attention *qua* agents, but since whatever is represented in attention is explicitly represented, it follows that for any PP agent *S*, if *S* has the capacity for attention, then *S* can explicitly represent itself *qua* agent. I make this argument in Sect. 3.2 before ending with some further remarks on theoretical import and suggestions for future research.

The structure of the paper is as follows. In Sect. 2, I distinguish self-representation from other ways of representing oneself. Next, after an overview of PP and SPPA, in Sect. 3 I defend (SNR). In Sect. 4, I address the objection that (SNR), rather than surprising, is actually trivial, and after a brief summary, in Sect 5. I discuss the upshot in greater detail and offer a brief suggestion for future research.

2 SELF-REPRESENTATION

In an influential paper, John Perry tells us a story.

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch (Perry, 1979, p. 3).

Perry's story illuminates an important distinction between two ways of representing oneself. On one hand, one can represent the person one just so happens to be, and this is Perry's initial predicament. For, he believes that a shopper has a torn sack and, unbeknownst to him, he just so happens to be that shopper. In some sense then, Perry believes that John Perry has a torn sack (call this belief Perry's "belief$_1$"). Importantly though, he doesn't believe, to borrow a phrase from Castañeda (2001), that John Perry *qua he himself* has a torn sack (call this belief Perry's belief$_2$). Not until he discovers that he is the shopper is Perry's belief of this latter kind. It is the latter kind of representation of oneself, what I will call "self-representation," in which I am presently interested (Perry, 1990).

Self-representations are distinguished by the role they play in the production of (visceromotor and skeletomotor) action. We can imagine, for example, that belief$_2$ will cause Perry to become red in the face with embarrassment as he frantically devises and executes a plan to clean up his mess. No such thing will happen with belief$_1$ alone, however, since belief$_1$ is simply not of the kind that plays this role.

What about self-representations, such as belief$_2$, is such that they play the role that they do? To answer this question, I'll need to introduce the concept of the *self-notion* (Perry, 1990). The self-notion is defined in terms of the relation in which it stands to two other important concepts in Perry's theory of self-knowledge. The first is what is called a *normally self-informative way of gaining information*:

> A perceptual state S is a normally self-informative way of knowing that f(y), when the fact that a person is in state S normally carries the information that that person is f, and normally does not carry the information that any other person is (Perry, 1990, p. 10).

Assuming that y ranges over all agents, seeing a coffee mug, for example, is a normally self-informative way of knowing (or more generally of gaining the information) that somebody is seeing a coffee mug because this perceptual state normally carries the information that the same agent (and not some other agent) who is seeing the coffee mug is seeing the coffee mug.
A normally self-informative way of gaining information is one that is "architecturally guaranteed—guaranteed by the way the organism is set up—to carry information about the agent to the same agent" (Van Leeuwen, 2012, p. 96).[4] For concision, I'll call information that an agent gains in normally self-informative ways *self$_N$-information*.[5]

[4] Perry calls it a "normally self-informative way of knowing," but for generality I called it a "normally self-informative way of gaining information." The qualification that these ways of knowing are "normally" self-informative excludes unusual cases in which, say, seeing a coffee mug is a way of knowing that somebody other than yourself is some such distance from that mug.

[5] Normally self-informative ways of knowing are immune to error through misidentification (Shoemaker, 1968), but they are not privileged nor are they incorrigible (Perry, 1990). Somebody else might be in a better position to have information that I can gain in a normally self-informative way, and information that I gain in a normally self-informative way, while unmistakably about me (immunity to error), may very well be wrong.

Just as important for the self-notion are what Perry calls “normally self-effecting actions.” An action is *normally self-effecting* if and only if that action (1) is guided by self$_N$-information, and (2) the agent that performs the action and the agent that is affected by the action are normally the same agent (Perry, 1990). Eating is a normally self-effecting action because it is one that is guided by self$_N$-information about the location of an agent’s body relative to its food, its mouth relative to its hands (if it has hands), and so on, and one by which the agent affected is normally just the agent performing the action (normally, only this agent gets fed).

Together, normally self-informative ways of gaining information and normally self-effecting ways of acting specify a repository of information: the repository of self$_N$-information that guides normally self-effecting actions. We call such a repository a *self-notion*.[6]

Self-notion: For an agent *S*, a mental representation *R* is a *self-notion* if and only if

(1) *R* only encodes self$_N$-information about *S* and
(2) that self$_N$-information guides *S*’s normally self-effecting actions.

With this in hand, we can now provide a definition of self-representation and answer the question we set out to answer at the beginning of this section.

Self-representation: A mental representation *R* is a *self-representation* if and only if

[6] Perry’s definition of the self-notion may appear to be circular. One might point out: the self-notion is partly defined in terms of normally self-effecting action, and normally self-effecting action is partly defined in terms of the self-notion. This apparent circularity is resolved, however, by the distinction I made earlier between the self-notion and the repository of self$_N$-information. These information repositories are not identical, but the former is part of the latter. Normally self-effecting actions are partly defined by being guided by the repository of self$_N$-information, *not* the self-notion. The self-notion is specified as the part of the repository of self$_N$-information that guides normally self-effecting action. In this way, normally self-effecting actions are defined independently of the self-notion so its definition is non-circular (Van Leeuwen, 2012).

(1) *R* is a constituent of *S*'s self-notion or
(2) *R* has *S*'s self-notion as a constituent.

What about self-representations, such as Perry's belief$_2$, is such that they play the role that they do? Self-representations play the role that they do because they involve the self-notion in some way. In particular, self-representations are either a constituent of the self-notion or they include the self-notion as a constituent (note that the self-notion is a self-representation in virtue of being a constituent of itself). For example, Perry's belief$_2$:

I am the shopper with the torn sack.

is a self-representation because it includes the self-notion as a constituent. It includes the self-notion as a constituent because it has *I* as a constituent, and, as Perry (1990) argues, the content of "I" is nothing but the content of the self-notion. Borrowing a convention from Van Leeuwen (2012), we can see this more clearly if we write

{Self-notion} am the shopper with the torn sack.

Perry becomes red in the face with embarrassment and moves his body to clean up his mess in virtue of his belief$_2$ which identifies him with the shopper. The self-notion guides normally self-effecting actions, but Perry's belief$_2$ identifies the content of his self-notion with that of his representation of the shopper, so not only his self$_N$-information, but also his information about the shopper will now guide his normally self-effecting actions. The fact that the shopper, who is identical to Perry, has a torn sack will guide Perry's actions in the store in a way that it didn't before (i.e., by causing him to become red in the face and to try to clean up).

Let's recap the developments of this section. First, one can have a representation of the agent one just so happens to be (e.g., Perry's belief$_1$). There is not much to say about this kind of representation other than that it is the kind one has of oneself prior to a self-discovery episode. In a self-discovery episode, one forms a representation of oneself that has one's self-notion as a constituent (e.g., Perry's belief$_2$). This brings us to a third and fourth kind of representation one can have of oneself. One can, first of all, have a self-notion, and when one predicates something of one's self-notion, one has a further kind of representation of oneself. Representations of this latter kind are often of great importance to us, for among them are beliefs like, "I am the shopper with the torn sack," but also those like "I am a philosopher," "I am a father," "I am a Christian," etc. Borrowing from Perry, I'll call representations of this kind "self-ideas."

What properties do these different kinds of representations have? We know that while self-notions and self-ideas are self-representations, representations of the agent one just so happens to be are not. Furthermore, unlike self-notions, self-ideas require the capacity for self-predication, that is, thoughts of the sort "I am *X*," which narrows the class of agents capable of forming self-ideas quite a bit. Self-predication may require, among other things, language. So it may turn out that only human beings (and perhaps some other mammals) are capable of forming self-ideas.

Many of our self-ideas are also explicit, by which I mean that many of our self-ideas explicitly encode information about ourselves (more on explicitness in Sect. 3.2). This shouldn't come as a surprise, for it is obvious that a representation such as

I am X.

explicitly encodes the information that I am *X*. What would be more surprising is if self-notions could encode information in this kind of way, for self$_N$-information, when first gained, is encoded implicitly in representations that explicitly encode information about a perceiver's environment. What I intend to show is just this. In particular, I will show that a recent framework for thinking about attention—the standard PP theory of attention—has the consequence that all agents endowed with a PP architecture and the capacity for attention are capable, not just of implicitly, but of explicitly encoding information about themselves in their self-notions.
This consequence is theoretically interesting because explicit self-representation is a building block for uniquely human self-ideas (e.g., "I am a thinking thing," "I am a Christian," etc.) which, for obvious reasons, are and have historically been of great philosophical, scientific, and cultural interest. If the standard PP theory of attention is correct, then most creatures in our world are a step closer than we might have thought (in the sense that they have this building block) to these self-ideas.

In the following section I provide an overview of PP and SPPA before defending (SNR).

3 ATTENTION AND SELF-REPRESENTATION IN PREDICTIVE PROCESSING

PP is a Bayesian model of cognition that was pioneered in neuroscience by Karl Friston and his colleagues (Friston et al., 2006; Friston & Stephan, 2007; Friston, 2009; 2010; Parr, Pezzulo, & Friston, 2022) and later developed in philosophy by Hohwy (2013), and Clark (2013, 2016).[7] PP's fundamental assumption is that PP agents (e.g., brain-bearing organisms) have access only to the effects of worldly causes, and in consequence must infer a probabilistic model of those causes. In particular, PP agents construct internal hierarchical generative models of the causes of their sensations. These models are *hierarchical* because they comprise ascending levels of prior hypotheses about causes in the world, where the higher levels constrain and predict the lower levels. They are *generative* in the sense that they explain how data are *generated*. For example, the hypothesis that it rains every Friday in Atlanta may constrain a lower-level hypothesis that says that it will rain this upcoming Friday. When Friday comes around, the internal model uses these hypotheses to explain why everything is wet. Why is everything wet? Because it is raining. Why is it raining? Because it rains every Friday in Atlanta.

PP agents aim to minimize discrepancies (prediction errors) between their internal models and actual sensory input, and they do so through perception (perceptual inference) and action (active inference). On the perception side, the agent updates its model (in an approximately Bayesian fashion) so that it better predicts sensory input. On the action side, the agent generates actions that selectively sculpt and sample sensory data that confirm its model. For example, if I expect there to be a mug on my desk when I walk into my office but there isn't one, then I can either update my beliefs about what objects are on my desk, or I can walk to the kitchen to get a mug to place on my desk.

[7] See Sprevak (2023) for a recent comprehensive review.

Under uncertain conditions, however, a discrepancy may not always be world-informative. In one context, the cause of movement in some bushes may be a lurking predator but in another only wind. From the perspective of the agent, however, the effect is exactly the same: the bushes are moving. It would be a fatal mistake, of course, to *always* infer predator or *always* infer wind. Thus, prediction error minimization must be sensitive to the context of prediction error so that only prediction errors that carry information about causal regularities in the world (world-informative prediction errors) get to update the internal model. To solve this problem, PP agents weight their prediction errors according to the expected precisions of those prediction errors in a process called *precision optimization*. This process selectively increases the gain on statistically precise (reliable) prediction error signals and attenuates imprecise (unreliable) ones so that only precise signals get to propagate up the perceptual hierarchy and update the internal model (Friston, 2009, 2010; Feldman & Friston, 2010). Since precise prediction error signals tend to be world-informative, it is world-informative prediction errors rather than noise that tend to update the internal model. It is thus by virtue of precision optimization that the PP agent can under uncertain conditions construct an accurate model that it can in turn use to successfully guide its actions in the world.

The standard PP theory of attention (SPPA) identifies attention with precision optimization (Friston, 2009, 2010; Feldman & Friston, 2010; Hohwy, 2013). More cleanly,

SPPA: Attention is precision optimization.[8]

[8] On this view of attention, attention is neither necessary nor sufficient for conscious perception (Hohwy, 2013). It is not necessary because conscious perception need not involve precision optimization since we can imagine a world that is sufficiently regular such that conscious agents do not need precision optimization in order to survive. Attention is not sufficient for conscious perception because precision optimization need not be conscious.

The appeal of SPPA lies in its simplicity, for according to the theory, attention is but a special form of prediction error minimization (in particular, because it is prediction error minimization on prediction errors, precision optimization is a second-order prediction error minimization). In Sect. 3.2, I will argue that if SPPA is correct, then all PP agents that have the capacity for attention can explicitly represent themselves *qua* agents. First, however, I will show that all PP agents, whether they have the capacity for attention or not, have self-models. Doing so will set us up for the main argument in Sect. 3.2.

3.1 From Self-Notions to Self-Models

Let's start by making the notion of the self-model precise.[9]

> **Self-Model**: A PP agent *S* has a *self-model* if and only if *S*'s internal model contains a *self-representation*.

Since self-notions are self-representations, having a self-notion would suffice (by definition) for having a self-model. I will argue that all PP agents have self-notions.

Since normally self-effecting actions are guided by self$_N$-information, and since the self-notion is the repository of self$_N$-information that guides normally self-effecting actions, if an agent can act in normally self-effecting ways, then that agent must have a self-notion. In order to show that agents act in normally self-effecting ways, we must show that agents act in ways that (1) are guided by self$_N$-information, and for which (2) the agent that performs the action and the agent

[9] Much of the existing literature on self-modeling concerns the underlying causes of the sense of self in the phenomenology of agency and perception. See, for example, Gallagher (2000), Metzinger (2003), Hohwy (2007), Pacherie (2006, 2007, 2008), de Vignemont (2011), Tsakiris (2011), Limanowski & Blankenburg (2013), and Frith (2015). What I would like to do in this section is show that all PP agents have self-models, regardless of whether they are capable of having conscious experiences.

that is affected by the action are normally the same agent. As I see it, there are two ways to do this. The first is to point to examples of normally self-effecting actions that all PP agents perform. Many simple actions such as eating, drinking, and scratching are normally self-effecting. When an agent eats or drinks, that agent is guided by self$_{N}$-information about the location of its body relative to its food or water, its mouth relative to its hands (if it has hands), and so on. Similarly, when an agent scratches, that agent is guided by self$_{N}$-information about the location of its itch relative to the limb or tree trunk that it will use to scratch itself. So, (1) agents that eat, drink, or scratch are agents that act in ways guided by self$_{N}$-information. Furthermore, (2) these actions normally do not lead any agent other than the agent performing the action to be fed, quenched, or scratched. So we have good reason to think that eating, drinking, and scratching are normally self-effecting actions, and if we accept that these actions are normally self-effecting, then we are already committed to saying that a whole swath of agents have self-notions since a whole swath of agents eat, drink, and scratch.

Having said this, there is another much stronger but perhaps more controversial way to show that all PP agents act in normally self-effecting ways. Consider the following claim.

> **All actions have normally self-effecting bases**: For every action, self-effecting or not, there exists a normally self-effecting sub-action on which that action depends.

If it turns out that all actions have normally self-effecting bases, then we will have much stronger support for the claim that all PP agents can act in normally self-effecting ways since by definition all PP agents have the capacity for action. While I will not defend this claim at length, I think I can show that it is at the very least plausible.

Consider the following examples. Contracting the muscles in your hand hard enough normally causes your hand to clench and not anybody else's. Contracting your biceps or your hamstrings hard enough normally causes your elbows or your knees to bend and not anybody else's. There is good reason to think that these actions are (or depend on) normally self-effecting ways of doing things (in the first case, of clenching a hand, in the second of bending an elbow, and in the third of bending a knee). For, (1) ordinarily, these kinds of actions are guided by proprioceptive feedback in processes of motor control (Frith, 2015; Wong, 2017), and proprioception is a normally self-informative way of gaining information since it is normally guaranteed to carry information about the agent to the agent. Additionally, (2) in performing these actions, you are normally the only agent affected (i.e., your hand clenches, not anybody else's, and your elbows or your knees bend, not anybody else's). So such actions are normally self-effecting. Now PP agents can minimize prediction error by acting on their environments, selectively sculpting and sampling data that confirm prior hypotheses. However, the selective sculpting and sampling of one's environment involves bodily movements that surely depend on simple actions of the sort that I've argued are normally self-effecting (e.g., muscular contractions). So it follows that all PP agents perform normally self-effecting actions. This of course means that all PP agents have self-notions, and since self-notions are self-representations, I conclude that all PP agents have self-models.[10]

[10] Perhaps the reader is not convinced by the foregoing argument. Doesn't my view entail that devices such as Roombas have self-representations? For, it seems that Roombas perform simple actions of the sort I've argued are self-effecting, but this would imply that they have self-notions which are a kind of self-representation. Yet, it would seem absurd to say that Roombas have self-representations. This is a good point, and I will give a more lengthy response to the more general objection (of which this is a special case) that the self-notion is too inclusive in Sect. 4. To summarize, while there are some reasons to think that the Roomba doesn't have a self-notion, it ultimately does not matter because the thesis is about the explicit self-notion which has more stringent requirements.

So all PP agents have self-models, and these self-models comprise, at the very least, self-notions, but in what do such notions consist? Well, we know that self-notions comprise information gained in normally self-informative ways, and that this information guides normally self-effecting actions, but in what do such ways of gaining information consist? I think there are a couple of good candidates. Interoceptive and proprioceptive signals, for example, reliably emit from the position of an agent's body in space and time. For this reason, we should expect that these information channels are normally guaranteed to carry information about the agent to the agent. This would mean that interoception and proprioception are normally self-informative ways of gaining information, and to the extent that the information that these channels relay guides normally self-effecting actions, interoceptive and proprioceptive priors, in other words, a primitive model of the agent's body, are good candidates for constituents of the self-notion. Now, it is one thing for an agent to have a self-model, as I argued all PP agents do, and another for an agent to *explicitly* represent itself *qua* agent. The latter is a more sophisticated cognitive achievement. Afterall, explicit self-representation is a building block of many more sophisticated forms of self-awareness (such as the capacity to think self-predicative I-thoughts). Having said this, I claim that PP agents explicitly represent themselves *qua* agents when they attend to themselves *qua* agents, and furthermore, all PP agents that have the capacity for attention can attend to themselves *qua* agents. We turn now to this issue.

3.2 From Self-Models to Self-Attention

Let's begin by distinguishing explicit from implicit representation. Consider how Dennett draws the distinction:

> … information is represented *explicitly* in a system if and only if there actually exists in the functionally relevant place in the system a physically structured object, a *formula* or *string* or *tokening* of some members of a system (or 'language') of elements for which there is a semantics or interpretation, and a provision (a mechanism of some sort) for reading or parsing the formula… for information to be represented *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly (Dennett, 1982-1983, p. 216).

We might summarize the foregoing definition as follows.

Explicitness: A representation *R explicitly* encodes information *I* if and only if

(1) *I* is physically realized in the system that has *R*,

(2) there exists a semantics (i.e., a system of meaning) for *I*, and

(3) there exists a decoding mechanism for *I*.

Implicitness: A representation *R implicitly* encodes information *I** if and only if

(1) *I** is not explicit,

(2) for some information *I, R* explicitly encodes *I*, and

(3) *I** is implied by *I*.

Consider the sentence-like mental representation expressed by “rocks are hard” that encodes the information *that rocks are hard*. We know that this representation encodes this information explicitly, for the information is physically realized as the sentence “rocks are hard” and there is a semantics (i.e., English language semantics) and a decoding mechanism (i.e., a language module) for that information. *That rocks are hard* also implies, however, *that some things are hard*, and thus “rocks are hard” implicitly rather than explicitly encodes the information *that some things are hard*. This information isn’t encoded explicitly in the system under consideration, however, since it has no physical realization in that system.

Importantly, explicit representation systems, as Dennett is sure to highlight, “need not be linear, sequential, sentence-like systems” (p. 216). Thus even if a given PP agent doesn’t have the capacity for sentence-like mental representation such as in the foregoing example, this would not alone guarantee that the agent does not have the capacity for explicit representation.

With this in hand, let’s consider explicit representation in PP agents. First, PP grants that the hierarchy of predictions that comprise an agent’s internal model is physically realized in that agent (e.g., as patterns of neural activity in the agent’s brain, or even, on more radical versions of PP, as the agent itself).[11] Furthermore, a semantics and decoding mechanism is provided for these predictions by the internal model itself. In particular, the hierarchical organization of predictions in the internal model is such that higher levels in the hierarchy predict (and in this sense give meaning to) lower levels (Parr, Pezzulo, & Friston, 2022). The prediction (in this case, the prior), for example, that it rains every Friday in Atlanta predicts a lower-level prediction that says that it will rain this upcoming Friday. When Friday comes around, the internal model uses these predictions to give meaning to current sensory input (there is water

[11] Some PP theorists, such as Friston, go as far as to say that PP agents *are* (that is, they are identical to) models of their environments. On this view, the model is physically realized as the PP agent itself.

falling from the sky and everything is wet). Why is everything wet? Because it is raining. Why is it raining? Because it rains every Friday in Atlanta.

Given these definitions, it follows from SPPA that representations in attention (i.e., high precision predictions) must always be explicit. Precision optimization selectively increases the gain on a high-precision prediction error, allowing that prediction error to propagate up the perceptual hierarchy, causing the agent to construct a revised high-precision prediction. This revised prediction is both physically realized in the agent and predicted by higher levels in the hierarchy, so it must explicitly represent what it predicts. This has the important consequence that an agent must explicitly represent *itself qua* agent in order to attend to itself *qua* agent. For concision, let’s call this special way of attending to oneself *self-attention.*

Self-Attention: *Self-attention* is attention to oneself *qua* agent.

Under SPPA an agent attends to an object if and only if that agent selectively increases the gain on a prediction error signal that carries information about a discrepancy involving a representation of that object in the perceptual hierarchy. Thus, an agent attends to itself *qua* agent, that is, it self-attends if and only if that agent selectively increases the gain on a prediction error signal that carries information about a discrepancy involving a self-representation (such as the self-model) in the perceptual hierarchy. Call this class of prediction error, *self-prediction error*. For example, an agent self-attends when that agent selectively increases the gain on a proprioceptive prediction error signal.

If we can show that all PP agents that have the capacity for attention (under SPPA) can selectively increase the gain on a self-prediction error signal, then we can show that all such PP agents can explicitly represent themselves *qua* agents. For the remainder of the paper, that is just what I intend to do.

My central argument proceeds as follows. First,

> (P1) For any PP agent *S*, if *S* has the capacity for attention, then *S* has a fallible self-model.

(P1) should be relatively straightforward at this point. In Sect. 3.1 I argued that all PP agents (whether they have the capacity for attention or not) have self-models. For, all PP agents have self-notions and having a self-notion suffices for having a self-model. Furthermore, self-notions are fallible since normally self-informative ways of gaining information are corrigible (Perry, 1990). To borrow an example from Perry, feeling that your face is flushed is a normally self-informative but corrigible way of knowing that you are blushing, for it is perfectly possible that you feel that your face is flushed but are wrong that you are blushing, in which case your self-notion is inaccurate (p. 9). So self-notions are fallible, and since having a self-notion suffices for having a self-model, the self-model is also fallible.

What is less obvious is that

> (P2) For any PP agent *S*, if *S* has the capacity for attention, then if *S* has a fallible self-model, then *S* can explicitly represent itself *qua* agent.

Let *S* be a PP agent that has the capacity for attention, and suppose that *S* has a fallible self-model. Since *S* has a fallible self-model, *S* is vulnerable to self-prediction errors. As an agent that has the capacity for attention, *S* must selectively increase the gain on high-precision prediction errors. This means that, in particular, *S* must selectively increase the gain on high-precision *self*-prediction errors. This shows that *S* can self-attend, which means that *S* can explicitly represent itself *qua* agent. Hence, for any PP agent *S* that has the capacity for attention, if *S* has a fallible self-model, then *S* can explicitly represent itself *qua* agent.

From (P1) and (P2) it follows that

> (SNR) For any PP agent *S*, if *S* has the capacity for attention, then *S* can explicitly represent itself *qua* agent.

So goes the argument.

It is worth speculating about what the contents of such explicit self-representations could be. Recall that in Sect. 3.1, I argued that a likely candidate for the self-notion is a primitive model of the agent's body (a set of interoceptive and proprioceptive priors). If this is true, then the agent is vulnerable to interoceptive and proprioceptive prediction errors, and the optimization of the precisions of these prediction errors constitutes attention to bodily signals such as pain, hunger, self-location, and so on. These are fine candidates for the contents of the explicit self-representations under discussion.

4 JUST HOW SURPRISING?

I've shown that (SNR): PP agents that have the capacity for attention can explicitly represent themselves *qua* agents. At the very beginning of the paper, I remarked that this consequence is surprising. But just how surprising is it? In this section, I would like to address the objection that (SNR) is not surprising. An opponent may defend this objection on the grounds that the self-notion is so inclusive that it renders (SNR) trivially true. Given how minimal the self-notion's requirements are, it is no surprise that all PP agents that have the capacity for attention have a self-notion that may, under the right conditions, become explicit.

I have several responses to this objection, but first, let me clarify exactly why I think (SNR) is surprising. Most of us likely have the intuition that most non-human organisms cannot represent themselves, or if they do, they only do so in a very weak sense. If non-human organisms can represent themselves, they can do so only accidentally (as in failed mirror recognition) or implicitly (by virtue of representing their environments). Intuitively, the self-representational capacities of non-human organisms end here. Dennett expresses this sentiment when he says,

> The hermit crab is designed in such a way as to see to it that it acquires a shell. Its organization, we might say, *implies* a shell, but the crab does not in any stronger sense *represent itself* as having a shell. It doesn't go in for self-representation (Dennett, 1991, p. 417).

I think many of us have this intuition. It also contradicts (SNR), and that is why I think (SNR) is surprising. If (SNR) is true, then explicit self-representation is *not* rare. Of course, that would not entail that Beetles and philosophers have the same self-representational capacities, just that Beetles and philosophers share one more building block for the self-representational capacities of philosophers than we might have thought, and I think the fact that predictive processing, or its standard theory of attention, has this consequence is interesting and worthy of the consideration of philosophers and scientists who have any sympathy for that theory (which right about now means many!).

With this clarification in hand, let's return to the objection. Is the self-notion too inclusive? Recall just what the self-notion is: it is what encodes all of the information an agent gains in normally self-informative ways that the agent also uses to guide its normally self-effecting actions. A system that has a self-notion thus must have two capacities: it must have the capacity for (1) normally self-informative ways of gaining information and (2) normally self-effecting action. Which systems have these capacities? For some systems, it is quite easy to tell whether they have them or not. A coffee mug that has on it printed the sentence "this coffee mug" in some very weak sense represents itself, but surely that mug does not have a self-notion, and for obvious reasons—for starters, it is not an agent. So, whatever the system is, if that system has a self-notion, it has to be an agent.

The forgoing point might be obvious, but I make it because it narrows the class of things that can have self-notions considerably. It rules out the coffee mug, but it also rules out, I presume, most computers (I do not think, for example, calculators have self-notions).[12]

That being said, which *agents* have capacities (1) and (2)? Consider normally self-informative ways of gaining information first. Human beings have this capacity, and a lot of other mammals probably do too. Perhaps you might think that the capacity for normally self-

[12] Which computers, if any, have self-notions is a technical question, and one that I am not qualified to answer.

informative ways of gaining information doesn’t extend beyond agents capable of conscious mental states (like pain). But supposing this were true, the range of animals who would then have the capacity for (1) is already quite impressive. Of course, however, the capacity for (1) *does* extend beyond agents capable of conscious mental states. All that is required for (1) is that an agent’s perceptual states normally carry information about that agent to that agent, and here is where things may get dicey for the skeptic. If all that is required for (1) is that the states of the agent normally carry information about that agent to that agent, it would seem that *all* perceiving agents have the capacity for (1) because it seems that percepts always carry information about the agent who has them to that very same agent. I will remain neutral on this point because whether or not it is true, (SNR) would still be surprising. For, if it is true that all perceiving agents have the capacity for (1), it does *not* follow that all perceiving agents have self-notions, for the self-notion in addition requires the capacity for (2) normally self-effecting action, which in turn narrows the class of agents that have self-notions quite a bit (more on this in the next paragraph). Thus, even if it is true that all perceiving agents have the capacity for (1), it does not follow that the self-notion is too inclusive, so (SNR) would still be surprising. Furthermore, if it is false that all perceiving agents have the capacity for (1), we can still generate an impressive (but not overly-inclusive) list of creatures that have the capacity for (1). Perhaps amoeba don’t gain information in normally self-informative ways, but dogs and maybe even fish still do.

A bit hangs on how much the requirement of the capacity for (2) narrows the class of agents that have self-notions. So, it is important to ask, which agents have the capacity for (2)? In Sect. 3.1, I argued that all agents, in virtue of their capacity for action, have the capacity for normally self-effecting action. While I think my argument for this claim shows that it is at least plausible, I recognize that it is controversial. But even if it is false, as I remarked in that section, it is still possible to give an impressive list of creatures who can act in normally self-effecting ways. Furthermore, as I argued in that section, such a list would also be a list of agents who have self-notions. So, how inclusive the concept of the self-notion is would depend on how inclusive such a list would be. But as with (1), it is not so straightforward to tell which agents have the capacity for (2). Since it is unclear which agents have these capacities, it is just as unclear which agents have self-notions. If one affirms that all agents have capacities (1) and (2), then one might have a case for the claim that the self-notion is too inclusive. But if one denies that all agents have these capacities, then it is at least unclear.

All that said, what I've so far done is try to persuade you that the self-notion is in fact not too inclusive. But actually, whether or not it is, it does not follow that (SNR) is not surprising. Recall, (SNR) is surprising because it contradicts the intuition that the self-representational capacities of non-human organisms are non-existent or very limited. If it is true that the self-notion is too inclusive, it does not follow that the *explicit* self-notion is too inclusive, for the requirements of the explicit self-notion are more stringent. Because these requirements are more stringent, and because the explicit self-notion is what matters, (SNR) remains surprising.

5 WHATEVER NEXT?

Let’s take stock. All PP agents, whether they have the capacity for attention or not, have self-models. For, they have self-notions, and having a self-notion suffices for having a self-model. Furthermore, the self-model inherits its fallibility from the self-notion, which is fallible since normally self-informative ways of gaining information are corrigible. As such, PP agents are vulnerable to self-prediction errors, that is, discrepancies involving self-representations (such as the self-model) in the perceptual hierarchy, and agents that have the capacity for attention (according to SPPA) must optimize the precisions of these prediction errors in self-attention. Crucially, self-attention is distinct from mere attention-to-self in that it requires that the agent represent itself in attention *qua* agent. This means that agents that have the capacity for attention can represent themselves in attention *qua* agents, but since representations in attention are explicit, it then follows that all PP agents that have the capacity for attention can explicitly represent themselves *qua* agents.

What is the upshot? If SPPA is correct, then everything from artificial PP agents and beetles all the way up to philosophers have the capacity for explicit self-representation. And given that the capacity for explicit self-representation is a building block for the more sophisticated forms of self-awareness in which philosophers and scientists alike have historically been most interested, this is theoretically important. It means that everything from artificial PP agents and lowly beetles all the way up to philosophers are a step closer to those more sophisticated forms of self-awareness than we might have thought. And depending on your philosophical sensibilities, you may find this conclusion to be absurd and a point against SPPA. On the other hand, you may find that, rather than absurd, it is actually quite modest. After all, the conclusion is only that these agents are more self-aware than we might have thought, but this does *not* entail that they can think about themselves in the ways that, say, humans think about themselves. I have only argued that these agents can have explicit self-notions, but as Perry notes, some agents have more sophisticated self-representational capacities (Perry, 2011). Human beings, for example, are not only able to learn about themselves in first-personal, normally self-informative ways, but also in third-personal, self-predicative ways. It is in this latter way that Perry comes to believe that he is the shopper with the torn sack. Many of the beliefs that we have about ourselves are, like this one, quite mundane, but many of them are also extremely important to us. Just consider a few:

I am a mother.

I am a friend.

I am a teacher

I am a good person.

I am American.

I am a Christian.

It is the capacity for beliefs such as these that puts the philosopher at a distance from the beetle. But while it is important to acknowledge this distance, it is also important not to overestimate it. What I hope to have shown is that this distance is an inch shorter than we might have thought.

Still, several questions remain, and among them are two kinds. The first is descriptive and internal to PP. As I clarified in the previous paragraph, even if everything from artificial PP agents and lowly beetles all the way up to philosophers have the capacity for explicit self-representation, it is not the case that these agents have the sophisticated self-representational capacities that are characteristic of creatures like us. There is much work to be done on the building blocks of those more sophisticated self-representational capacities in PP agents.

The second kind of question is normative and external to PP: How should the conclusion that everything from artificial PP agents and lowly beetles all the way up to philosophers have the capacity for explicit self-representation guide the philosophical and scientific explication of attention and self-representation? If I am right, then (1) our intuitions about self-representation, (2) SPPA, (3) Perry's theory of self-knowledge, and (4) PP are incompatible. That is, at least one must go. While a well-informed theoretical choice will require a more careful analysis of the relative virtues of (1)-(4), I would nonetheless like to end with a suggestion.

Most of the problems that come with radical PP result from its attempt to reduce everything every organism does to prediction and prediction error minimization.[13] Perhaps, however, prediction error minimization is only something brains do, or only something *some* brains do, and maybe not *all* those brains ever do.[14] (SNR) bears on all PP agents that have the capacity for attention, but which agents are PP agents? If they are only a limited number of agents with brains, then we get to keep our intuitions, SPPA, and Perry's theory. The cost, of course, would be what makes PP most attractive: its sweeping scope. But, perhaps a more modest theory is just what we need.

[13] See Friston, Thornton, and Clark (2012), Hohwy (2013), Clark (2017), Klein (2018), and Sun and Firestone (2020) for discussion.

[14] In this way, this paper serves as a contribution to the debate about the scope of PP (Sims, 2017).

REFERENCES

Bowman, H., Collins, D. J., Nayak, A. K., & Cruse, D (2023). Is predictive coding falsifiable? *Neuroscience & Biobehavioral Reviews*, *154*, 1-30. https://doi.org/10.1016/j.neubiorev.2023.105404

Castañeda (2001). 'He': A study in the logic of self-consciousness. In A. Brook & R. C. DeVidi (Eds.), *Self-reference and self-awareness* (pp. 51-80). John Benjamins Publishing Co.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-253. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Clark, A. (2017). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*, 521–534. https://doi.org/10.1007/s11097-017-9525-z

Dennett, D. C. (1982-1983). Styles of Mental Representation. *Proceedings of the Aristotelian Society, 83*, 213-226.

Dennett, D. C. (2017). *Consciousness Explained.* Little, Brown and Co.

Feldman & Friston (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, Article 215. https://doi.org/10.3389/fnhum.2010.00215

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris, 100*(1-3), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K. J. & Stephan, K. E (2007). Free-energy and the brain. *Synthese, 159*(3), 417-458. https://doi.org/10.1007/s11229-007-9237-y

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293-301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11, 127-138. https://doi.org/10.1038/nrn2787

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, Article 130. https://doi.org/10.3389/fpsyg.2012.00130

Frith, C. D. (2015). *The cognitive neuropsychology of schizophrenia* (classic ed.). Psychology Press. https://doi.org/10.4324/9781315749174

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, *4*(1), 14-21. https://doi.org/10.1016/S1364-6613(99)01417-5

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche, 13(1)*, 1-20.

Hohwy, J. (2013). *The predictive mind.* Oxford University Press.

Klein, C. (2018). What do predictive coders want? Synthese, 195, 2541–2557. https://doi.org/10.1007/s11229-016-1250-6

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free-energy principle. *Frontiers in Human Neuroscience*, *7*, 1-12. https://doi.org/10.3389/fnhum.2013.00547

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*, 353–393. https://doi.org/10.1023/B:PHEN.0000007366.42918.eb

Pacherie, E. (2006). Towards a dynamic theory of intentions. In S. Pockett, W.P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior?: An investigation of the nature of volition* (pp. 145-167). MIT Press.

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche: An Interdisciplinary Journal of Research on Consciousness, 13*, 1–30.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*(1), 179–217. https://doi.org/10.1016/j.cognition.2007.09.003

Parr, Pezzulo, & Friston (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Perry, J. (1990). Self-notions. *Logos*, 17-31.

Perry, J. (2011). On knowing one's self. In S. Gallagher (Ed.), *The Oxford Handbook of the Self* (pp. 372–393). Oxford.

Ransom, M. & Fazelpour, S (2020). The many faces of attention: Why precision optimization is not attention. In D. Mendonça, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 119-139). Bloomsbury Publishing.

Shoemaker, S. (1968). Self-Reference and Self-Awareness. *The Journal of Philosophy, 65*(19), 555-567.

Sims, A. (2017). The problems with prediction: The dark room problem and the scope dispute. In T. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing.* Mind Group. https://doi.org/10.15502/9783958573246

Sprevak, M. (2023). Predictive coding I: Introduction. *Philosophy Compass,* e12950. https://doi.org/10.1111/phc3.12950

Sun, Z. & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences, 24*(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006

Tsakiris, M. (2011). The sense of body ownership. In S. Gallagher (Ed.), *The Oxford Handbook of the Self* (pp. 180-203). Oxford.

Van Leeuwen (2012). Perry on self-knowledge. In Newen, A. & van Riel, R. (Eds.), *Identity, language, and mind: An introduction to the philosophy of John Perry* (pp. 89-107). CSLI Publications.

de Vignemont, F. & Fourneret, P. (2004). The sense of agency: a philosophical and empirical review of the "who" system. *Consciousness and Cognition, 13*, 1-19. https://doi.org/10.1016/S1053-8100(03)00022-9

Wong, H. (2017). On proprioception in action: Multimodality versus deafferentation. *Mind and Language, 32*(3), 259-282. https://doi.org/10.1111/mila.12142