Topoi (2020) 39:559–574
https://doi.org/10.1007/s11245-018-9561-5

Free Energy and the Self: An Ecological–Enactive Interpretation

Julian Kiverstein[1]

Published online: 12 April 2018

Abstract
According to the free energy principle all living systems aim to minimise free energy in their sensory exchanges with the environment. Processes of free energy minimisation are thus ubiquitous in the biological world. Indeed it has been argued that even plants engage in free energy minimisation. Not all living things however *feel* alive. How then did the feeling of being alive get started? In line with the arguments of the phenomenologists, I will claim that every feeling must be felt by someone. It must have mineness built into it if it is to feel a particular way. The question I take up in this paper asks how mineness might have arisen out of processes of free energy minimisation, given that many systems that keep themselves alive lack mineness. The hypothesis I develop in this paper is that the life of an organism can be seen as an inferential process. Every living system embodies a probability distribution conditioned on a model of the sensory, physiological, and morphological states that are highly probably given the life it leads and the niche it inhabits. I argue for an ecological and enactive interpretation of free energy. I show how once the life of an organism reaches a certain level of complexity mineness emerges as an intrinsic part of the process of life itself.

Keywords Free energy principle · Mineness · Minimal self · Ecological enactive · Active inference · Relational self · Multisensory integration · Bayesian brain

1 Introduction

Every conscious experience is an experience for someone, the self or subject of this experience. Feelings are felt by someone, thoughts are thought by someone, and experiences are experienced by someone. Self-consciousness is not some extra property added to consciousness but is *intrinsic* to it. It is a part of its very mode of being of conscious experience that it presents the world as appearing a certain way *for* someone, for me or you (Sartre 2003, p. 100). Self-consciousness is not the outcome of introspection or reflection. It is not something that only occurs under exceptional circumstances, when a person deliberately makes a conscious episode they are undergoing into the object of their attention. Instead self-consciousness "is a feature characterising the experiential dimension as such, no matter what worldly entities we might otherwise be intentionally directed at" (Zahavi 2014, p. 27). Self-consciousness is an invariant, always present, ubiquitous feature of our consciousness of the world. It is in other words a part of the phenomenological structure of conscious experience.

This basic form of self-awareness I will henceforth refer to as "mineness". It refers to the feature of experience whereby every experience is an experience for a self. Without this feature of mineness there would be nothing it is like to perceive. Perception would lack phenomenality.[1] The perceptual episode wouldn't make anything in the world manifest. Sensations wouldn't feel like anything because they wouldn't be sensed by anyone. It is only because feelings are felt, experiences are experienced and so on that these episodes present the world as appearing or feeling a certain way. Experiences owe their very phenomenality—their making things manifest to someone—in part to their mineness.

The view of mineness I will develop in this paper starts from the following three interrelated theses:

✉ Julian Kiverstein
j.d.kiverstein@amc.uva.nl

[1] Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

[1] I first encountered this account of phenomenal experience in Dan Zahavi's writings—see e.g. Zahavi (2005, 2014, 2017). Zahavi shows how his use of the term "mineness" to characterise phenomenal experience can be traced back to the writings of the phenomenological philosophers at the beginning of the twentieth century. See also Gallagher (2000, 2005), Legrand (2006, 2012), Thompson (2007).

1. Phenomenal experiences owe their phenomenality to their being episodes that occur for someone, a subject or self that is self-aware.
2. "Mineness" is a form of basic or minimal self-awareness intrinsic to phenomenal experiences.
3. Every experience has minimal self-awareness built into it as a part of its phenomenological or intentional structure. Experiences can only present the world as appearing a certain way because they are experiences for a self, me or you.

I will henceforth refer to this view as the "phenomenological theory of selfhood" (abbreviated to "the phenomenological theory"). The phenomenological theory implies that any naturalistic explanation of phenomenal consciousness in the terms of psychology and the neurosciences will need to explain how mineness can be intrinsic to phenomenal consciousness. Minimal self-awareness is a part of what makes a mental state a *phenomenally conscious* mental state. Any naturalistic explanation of consciousness must include an explanation of the intimate relationship between being a self and being phenomenally conscious of the world.

In what follows I will propose such a naturalistic explanation that takes selfhood to emerge out of self-organising biological processes. Organisms maintain their organisation in their sensory exchanges with the environment. They seem to resist a tendency to disorder that otherwise applies to physical systems more generally (Friston and Stephan 2007). They do not randomly explore the space of possible sensory and physiological states. Instead they regulate their interactions with the environment so as to ensure that they visit and revisit a limited range of sensory and physiological states (Friston 2010). They keep themselves in the sensory and physiological states that define the conditions of their existence. Hearts beat rhythmically, body temperature tends to stay within certain bounds, animals engage in regular routines and habitual behaviours. When these regular rhythms of life are disturbed (for instance by the sudden appearance of a predator leading to an increase in heart rate and breathing), the organism takes measures to return their bodily processes to the highly probable, regular rhythms that are expected. They engage in fight or flight (Friston 2017).

The hypothesis I develop in this paper is that the life of an organism can be seen as an inferential process. Every living system embodies a probability distribution conditioned on a model of the sensory, physiological, and morphological states that are highly probably given the life it leads and the niche it inhabits. The organism embodies in its biological organisation a hierarchically structured model of its *own* existence in its environment, or equivalently its being-in-the-world. Consider for instance how dancers come to embody in their muscles the activities they repeatedly engage in as dancers.[2] As a model of its own existence, the organism can accurately predict the sensory consequences of its own interactions with the environment over multiple temporal and spatial scales. In this way, the organism will be able to maintain its organisation in its sensory and energetic exchanges with the environment, keeping itself in the narrow range of sensory and physiological states it needs to occupy given the kind of life it leads. However, in a constantly changing environment replete with sensory noise, what the organism expects to happen in its sensory encounters with the environment will often not come to fruition. The predictions it makes will fail to match its incoming flow of sensory information, and the organism will need to adapt its organisation accordingly. Thus, in order to maintain its organisation in its sensory exchanges with the environment, the organism will need to keep the discrepancy between the predictions of its model and what actually ensues to a minimum, or what is technically referred to as "prediction error".

Processes of prediction-error minimisation are however ubiquitous in the biological world. Indeed it has been argued that even plants engage in prediction-error minimisation (Calvo and Friston 2017). Not all living things however *feel* alive. Plants almost certainly don't. Bacteria are capable of a minimal form of purposive agency (Fulda 2017; Di Paolo et al. 2017). Since they lack a brain for regulating the changing internal states of their bodies, they probably also lack feeling (Thompson 2015, pp. 335–336). What about insects that continue behaving in the same way when they have suffered severe injuries? They too most likely lack feeling insofar as their behaviour seems to be unaffected by damage to their bodies. However, it is an assumption of this paper (to be unpacked further in Sect. 2 below) that all living beings engage in prediction-error minimisation. For as I have already indicated in the introduction, there is an intimate connection between a living system's ability to maintain its own organisation (a process known as "autopoiesis") and prediction error minimisation.[3] But how then did the feeling of being alive get started?

This is of course among the central puzzles in the science of consciousness, and no one yet has the full solution. In line with the arguments of the phenomenologists, I've claimed that every feeling must be felt by someone. It must have mineness built into it if it is to feel a particular way. The question I take up in this paper asks how mineness might have arisen out of processes of prediction-error

[2] My thanks to Erik Rietveld for this example.

[3] For more discussion of the connection between autopoiesis and prediction-error minimisation (see Allen and Friston 2017; Kirchhoff 2017; Kirchhoff and Froese 2017; Bruineberg et al. 2017).

minimisation, given that many systems that keep themselves alive lack mineness?[4]

In Sect. 1 I take up this question by speculating about how subjectivity might have emerged early in evolutionary history out of processes of action control. I show how subjectivity might be thought of as tied to processes of purposive agency on the one hand, and sensorimotor integration on the other. Neither of these ingredients suffices individually for subjectivity but when they work together as a single cognitive package I suggest the result might plausibly be the beginnings of a subjective mental life. The key question is how these two capacities might have come to be combined. I suggest the connection between life and probabilistic inference might provide the answer to this question. In Sect. 2 I explain why it might make sense to understand life in terms of inference. I introduce the free energy principle, according to which any living system will aim to minimise free energy in its sensory exchanges with the environment. Section 3 shows how such a process of free energy minimisation is accomplished through a process referred to as "active inference". I explain how active inference is best understood in ecological and enactive terms. The sensory and physiological states the organism expects to occupy relate to its way of life in its niche, and the affordances it provides for animals with this way of life. Prediction-error should therefore be understood in terms of the coupled dynamics of the animal in its eco-niche that lead towards dynamic equilibrium (Bruineberg and Rietveld 2014; Bruineberg et al. 2017). Active inference doesn't however suffice to explain mineness, since every living system will keep its own free energy to a minimum through a process of active inference. Section 4 therefore takes up the question of what would need to be added to active inference to yield mineness. I look to recent theoretical work by Karl Friston for an answer to this question (Friston 2017, 2018). It follows from Friston's arguments that once an organism reaches the level of complexity so that it can act to minimise its own *expected* free energy, mineness will emerge as intrinsic to the process of life itself. Section 5 closes the paper by engaging with recent attempts that have been made in the cognitive neuroscience literature to understand the self in terms of active inference. I argue that if these proposals are to explain mineness, they are best interpreted in ecological and enactive terms. I close by showing how the resulting account of mineness supports a relational theory of the self. This is because on this ecological and enactive reading of active inference the organism and its environment are co-specifying, and co-determining.[5] The argument of this paper is that any organism capable of action control of the right complexity will also be a self. Since organisms are best understood in relation to their environment, so also are selves.[6]

2 The Origin of Feeling

In his famous paper on the reflex arc, John Dewey described how perception and action form a circuit in which "the motor response determines the stimulus, just as truly as sensory stimulus determines movement" (Dewey 1896). Motor behaviour determines sensation in the sense of controlling or regulating the sensory information the perceptual systems detect. The activity of the organism determines the sensory information that flows back into the brain as it purposefully engages with its environment. What I want to take from Dewey is his emphasis on the *active* control of the flow of sensory information. In what follows I will explore the possibility that mineness may have its origins in processes of action control.[7] Feeling relates to action possibilities provided by the environment that matter to the organism. Think here of inner feelings of pleasure and pain as core examples of what I have in mind. The occurrence of these episodes literally demands that the organism do something. They inform the organism that something important is happening

[4] Such a hypothesis has been developed in the cognitive neuroscience literature in important papers by Seth (2013), Apps and Tsakiris (2014), Limanowski and Blankenburg (2013). I draw on their arguments in what follows but develop their ideas in a somewhat different direction in defending an ecological and enactive interpretation of the free energy principle.

[5] Gibson explains this co-determination of the animal and its environment well when he writes: "The fact is worth remembering because it is often neglected that the words *animal* and *environment* make an inseparable pair. Each term implies the other. No animal could exist without an environment surrounding it. Equally, although not so obvious, an environment implies an animal (or at least an organism) to be surrounded" (Gibson 1979, p. 4).

[6] My account differs in some respects from recent work on the relational self that take the self to be relationally constituted because of intersubjective relations between infants and caregivers. Such intersubjective relations have been shown to play a crucial role in shaping early experiences of being embodied through for instance experiences of affective touch (Fotopoulou and Tsakiris 2017; Ciaunica and Fotopoulou 2017). I briefly discuss the relation of my account of mineness to these relational theories in my concluding comments.

[7] See Godfrey-Smith (2016, p. 96) for a related proposal. An anonymous reviewer reminded me that the seeds of this idea were also present in first generation Cybernetics and in the control theory of perception (Powers 1973). The central idea of the control theory of perception is articulated above—it is that action is for the control of perception rather than the other way around. (See also Anderson 2014, Chap. 5). In a different context in discussing the porousness of the boundaries of the self, Clark (2003) also argues for a conceptual connection between being a self and action control. Clark's interest is with the question of what makes it the case that something that lies outside of the body can nevertheless fall within the boundaries that demarcate the self from the rest of the world. The answer Clark returns to this questions appeals to the control the person has over an external resource.

and that taking appropriate action is a matter of imperative (Klein 2015).

Feeling I suggest, is tied to an animal's practical engagement with its environment. As the nervous systems of organism grew in complexity, so also did the repertoire of action possibilities available to them. Their subjective mental life gradually underwent transformation. I am thus in agreement with what Peter Godfrey-Smith calls the "transformation view" of subjective experience. The transformation view holds that subjective experience emerged in evolutionary history before such late-emerging cognitive processes as working memory, global workspaces and multisensory integration. These cognitive functions may be implicated in the types of subjective experience humans enjoy, but they should be thought of as transformations and complexifications of the types of subjective experience found in more simple lifeforms much earlier in time.[8]

The capacity for purposive agency may be necessary for subjective experience but it isn't sufficient. Being selectively and differentially moved to respond will only yield feeling or subjective experience if the bodily states that move the organism are bodily states *for* the organism. This is to say they must have "mineness". It is here that sensory feedback gets to do important work. Mineness has been hypothesised to have its biological roots in the integration of sensorimotor efference and reafferent feedback (Christoff et al. 2011; Thompson 2015). As the organism is moved to engage with the relevant affordances of her environment, her movements are met with sensory feedback in the form of proprioception, kinesthesis, vision and other forms of exteroception. There are systematic, predictable relationships that hold between movement and the sensory feedback that is the consequence of those movements. This anchoring of movements in sensory feedback has been argued to be "self-specifying" (Christoff et al. 2011). Thus to rehearse an example of Christoff and colleagues, when I bite into a lemon and experience the sour taste of the lemon, my activity of biting has systematic effects on my olfactory and gustatory senses—I taste and smell the lemon. I can see the lemon in my hand, and I can feel my teeth biting into it. Processes of sensorimotor integration in which action is integrated with its sensory effects are "self-specifying" because the sensory feedback that is integrated with motor behaviour arises from the organism's own activities. Sensorimotor integration thus forms the biological basis for "a unique egocentric perspective in perception and action." (Christoff et al. 2011, p. 106).

Processes of homeostasis are also self-specifying. Homeostasis refers to the self-regulatory processes that ensure the internal states of the organism that relate to the preservation of its life remain within a tightly restricted range of possible values. Think of the maintenance of blood sugar levels or of bodily temperature. Regulation will often come in the form of the activities of the organism—for example, finding a blanket or lighting a fire to warm the body when returning home on a cold winter's day. These activities have systematic effects on the internal conditions of the body (e.g. bodily temperature), and tracking such systematic relationship between efference (action) and reafferance (sensory effects of action) is a self-specifying process. It supports "an implicit feeling of the body's internal condition in perception and action" (Christoff et al. 2011, p. 106).

Now I suggest that neither sensorimotor integration nor purposive agency suffices for subjective experience when considered as separate processes that function in isolation. However, consider what might happen when these processes are considered as working together as parts of single package? Purposive agency is driven by the organism's concern to maintain its own viability in its practical engagement with the environment. Basic drives will include bodily needs that relate to temperature, hunger, and so on. Now combine actions that are driven by such basic needs with a capacity for tracking systematic relationships between perceiving and acting. Cycles of perception and action will systematically relate efference to reafferance so as to establish a point of view on the world. These cycles of perception and action will be motivated by the organism's concern for how it's fairing in the world. The environment will thus show up as offering action possibilities *for* the organism that matter to the organism in some way. I suggest the organism will then qualify as a subject of experience.

A key assumption of this paper is that mineness may have first emerged in the physical world when lifeforms began to control and regulate their activities based on what they care about as living beings. How might sensorimotor integration and purposive agency be combined so as to work together for an organism? It is in answering this question that I suggest it might be helpful to think about living systems as modelling their own existence in an attempt to continuously keep their own free energy to a minimum. In the next section I explain why it might make sense to understand life in terms of inference. This will allow us to get the core idea of free energy minimisation in play.

[8] Exactly when in the history of life feeling first emerged needn't concern us further. Perhaps bacteria are capable of some minimal degree of feeling to the extent that they are capable of purposive agency (see Fulda 2017). Godfrey-Smith speculates simple experience may have begun in the Cambrian era that saw an explosion of lifeforms capable of "richer forms of engagement with the world". Ediacaran animals he suggests may have had nervous systems that were primarily taken up with sensorimotor coordination which they could achieve without much in the way of control (Godfrey-Smith 2016, pp. 96–97).

3 The Free Energy Principle Introduced

In my introduction I suggested that organisms can be thought of as embodying in their biological organisation a model of their own existence. The organism is defined by the bodily states that it needs to maintain within certain bounds if it is to continue to exist. For example, the core body temperature of humans fluctuates around 37 °C. Such a body temperature is on average and over time highly probable since it is the temperature the human body needs to maintain at a constant level if it is to remain healthy. Deviate too far from such an equilibrium (by for instance remaining outdoors in a sub-zero environment for too long), and this can prove dangerous. The agent must take immediate measures to ensure their body temperature returns to its local equilibrium. There is thus a set of internal bodily states (exteroceptive and interoceptive) that the organism *expects* given its phenotype and the eco-niche it lives in, or what I will henceforth refer to as its existence (or way of life). The bodily states the organism expects are the bodily states that are highly probable given its existence. The organism's expectations "model" its existence in the sense of explaining or accounting for its continued existence. The organism continues to exist by ensuring that it remains in bodily states that are highly probable on average and over time and avoids bodily states that it has a low-probability of finding itself in because were it to regularly frequent such states over time it would cease to exist.

The organism thus embodies in its biological organisation a model of its own existence. The organism is a model of itself in its eco-niche, it is what I will call a "self-model", where the organism should be understood in relation to its niche. The organism and its environment form a complementary pair (Gibson 1979; Bruineberg et al. 2017). As I will use the term, a "self-model" is not yet a model of a self.[9] All organisms are self-models in my sense of the term because they have a biological organisation that models their own existence. However, as we've just seen in Sect. 1, not all organisms are selves. What needs to be added to a self-model for an organism to qualify as a self is mineness. I've suggested the special ingredient that yields mineness might be a cognitive package that combines purposive agency with sensorimotor integration. My aim in this paper is to explicate how a self-model might come to generate mineness. This I will do by using the theory of the organism as a self-model to explain how purposive agency and sensorimotor integration might be combined.

Let us begin first by clarifying why it makes sense to think of the organism as having a biological organisation that accounts for or explains its own existence. We've seen how the internal bodily states the organism *expects* are those it must actively maintain within a tightly restricted range of values as long as the organism succeeds in sustaining its way of life (Friston and Stephan 2007; Friston 2010). They are the possible bodily states the organism must visit and repeatedly revisit if it is to maintain the way of life it embodies. The organism's existence can thus be defined as a probability density conditioned on a "self-model" which identifies the bodily states it is highly likely to find the organism occupying on average and in the long run given its way of life (Hohwy 2017, p. 2). Life is a process of inferring a model that identifies the bodily states the organism has a high probability of finding itself in given its way of life. When the organism occupies such highly probable bodily states, it can thus be said to *maximise* the evidence for the self-model it has inferred as the explanation of its own existence. Bodily states count as evidence for a self-model when the self-model can explain or account for them. Highly probable bodily states *maximise* the likelihood of a self-model. They do so for the following reason. As long as the organism behaves so as to keep itself in the internal bodily states it expects to be in, the organism will count as having a biological organisation that stays well-adapted to its environment. Its internal bodily states will maximise the likelihood that its biological organisation is a good self-model, a model of its existence that keeps the organism well-adapted to its niche.[10]

Now consider what happens when the organism finds itself occupying bodily states that fail to provide evidence for the self-model it embodies. These are internal bodily states that are unlikely or surprising given its continued existence. If the organism were to repeatedly visit such states over the long run this would threaten its viability, and may in the end lead to death. The term "surprise" is being used here in an information-theoretic sense. It is the negative log probability associated with a bodily state. Surprise increases as a function of the improbability of the organism finding itself in such a bodily state. The average surprise over time is entropy or disorder (Friston 2012). Bodily states that are surprising have a low probability of occurring over time because an organism that regularly frequented them would be dispersed nearly everywhere leading to its own untimely

[9] Here I depart from standard terminology as employed in Metzinger (2003, 2009) for instance. My non-standard use of the term has the virtue of keeping the door open for non-representational accounts of the self, which I will eventually argue will prove necessary if we are to explain mineness in terms of processes of free energy minimisation.

[10] Natural selection can thus be cast as a process of self-model selection (Campbell 2016; Ramstead et al. 2017). The organism's phenotype can be thought of as a self-model, and natural selection as the process of selecting self-models with the greatest evidence (i.e. that best account for the bodily states the organism occupies on average and in the long run). The self-model that is maximally supported by evidence is the self-model that does the best job of adapting the organism to its environment.

demise, its disintegration, dissipation and dispersal into the environment.

The organism however has no direct way of knowing which of its bodily states it should expect to occupy on average, and in the long run. To know this it would have to somehow be able to evaluate an intractable number of possible states of being, so as to assess of its possible states of being which it is most likely to occupy on average and over time. The way organisms get around this otherwise intractable problem is by using what is referred to as "free energy" a knowable quantity that can be used to bound surprise, the probability of finding itself in low-probability bodily states. This is a trick the organism can depend upon because free energy can be shown mathematically to be an upper-bound on surprise (i.e. free energy is always greater than surprise) (Friston 2010). Thus so long as the organism keeps free energy to a minimum the organism will also succeed in avoiding occupying surprising (or low-probability) bodily states. But what is free energy?

Thermodynamic free energy refers to the energy available in a physical system (for instance a gas) that can be put to useful work. For example, consider a collection of gas molecules with the same kinetic energy. Only part of this energy is available to do work. This depends upon the entropy or 'orderliness' of molecular motion. For example, a low entropy system would correspond to molecules all moving in one direction (i.e., like a wind). This means that nearly all the energy is available to do work (i.e., turn a turbine). Conversely, if the entropy is high—and all the molecules are moving in random directions—there is no useful work to be harnessed.[11] Free energy correlates with unpredictability. The more unpredictable the location of particles, the more free energy there is overall in the gas to be put to work. This is because the behaviour of the particles in the gas is relatively disorderly. Conversely the more predictable the position of the particles, the more certain we can be of a particle's location at any given time. The energy available within the gas is not being used to disperse particles across space randomly, leading to unpredictability, but is instead already being put to work.

The notion of "free energy" I will employ in this paper is however not thermodynamic free energy as it applies to for instance the behaviour of the particles in a gas. It is instead *informational* free energy, a measure of the probability of bodily states conditioned on a self-model. In what follows I will be using the term "free energy" in this information-theoretic sense (Tribus 1961; Friston 2010).

Free energy as I will employ the term is the useful work a self-model does in predicting an organism's bodily states in its sensory exchanges with the environment. The better the self-model performs at predicting the organism's bodily states, the lower the free energy associated with this model. Reducing free energy thus amounts to improving the fit of the self-model to the data—the organism's bodily states that arise in its sensory exchanges with the environment. If the organism is to maintain its own viability, it should aim to keep free energy to a minimum. It should aim to be a self-model that succeeds in predicting its own bodily states, or equivalently that keeps the discrepancy between its predicted and current bodily states to a minimum. In doing so it will ensure that it remains in bodily states that are likely given the kind of life it leads, and avoids finding itself in bodily states that are unlikely because they are a threat to its way of life.

The free energy principle states that all quantities that can change in living organisms will change in such a way as to minimise free energy expected under a particular course of action. The organism's adaptive behaviour is the process of optimising the parameters of its self-model so that the organism does better and better at adapting to the random fluctuations of its environment. Over time the organism aims to keep to a minimum the discrepancies between the predictions of its self-model and the bodily states it finds itself occupying as a consequence of interacting with the environment. It does so through a process of optimising the parameters of its model that serve as the basis for its predictions. In this way the organism ensures that it has an internal global dynamics that allow it to stay well-adapted to its environment. In the next section I explain how this process of optimisation works through introducing the concept of active inference. I show how active inference is best understood in ecological and enactive terms. This will then provide us with all the conceptual tools needed to explain how purposive agency and sensorimotor integration can be combined in a free energy minimising system so as to deliver mineness.

4 Active Inference: The Ecological and Enactive Interpretation

It is through active inference that the organism keeps free energy to a minimum in its interactions with the environment. We've seen how the living system is a self-model—it is defined by the bodily states that on average and over time it has a high probability of instantiating. In its interactions with the environment the organism samples the world so as to gather evidence that maximizes the likelihood of the self-model it has become. Action is understood here as the organism's means of reducing free energy—the organism predicts the bodily states it expects to occupy based on its self-model and then acts so to fulfill its predictions. Suppose for instance you are feeling hungry. This feeling arises

[11] I thank an anonymous reviewer for help with the notion of thermodynamic free energy and its relation to variational free energy.

because your body predicts being well-nourished—this is among the highly probable bodily states your body is fortunate to be able to expect. It therefore registers a mismatch between its predictions and the body's current energy levels. This prediction is held constant and not updated because the prediction error is treated by the body as one it ought to pay attention to.[12] The ensuing actions serve to reduce this discrepancy through finding and eating some food. It exploits opportunities the environment offers for restoring your body's energy levels to their expected equilibrium state.

The self-model can thus be thought of as regulating the coupling of the agent to the environment through perception action cycles, and doing so in such a way as to (on average and over time) keep free energy to a minimum. What needs to be accounted for by the self-model is the generation of observations in the form of changing exteroceptive and interoceptive bodily states, (or what I will continue to abbreviate as "bodily states"). Such an account is inferred through learning statistical regularities that map bodily states onto actions and hidden environmental causes. The self-model is a model of causal and statistical structure that maps relations of dependence between states of the environment, actions of the organism and the sensory consequences of those actions. But it is through cycles of perception and action that such observations are generated. Thus the self-model should be thought of as the outcome of a process of inferring the causal dynamics of agent–environment interactions.

The self-model is not an *agent-neutral* model of the hidden environmental causes of sensory input. It is not a model of the hidden causal structure in the environment of the sort a scientist might construct (Bruineberg et al. 2017). It is a model of patterns of agent–environment interactions that are most likely to minimise free energy. Sensory input is generated by an agent's actions performed in a particular situation as the agent acts upon possibilities for action provided by its environment. Thus it is on the basis of the actions the agent performs that inferences are made about the hidden causes of sensory input. It is the environmental causes of sensory input in relation to the organism's own actions that is being inferred by the generative model.

Affordances are possibilities for action the environment offers to agents with the necessary abilities. Thus putting all of this together we can say the self-model is a model of the agent's selective engagement with affordances (Bruineberg and Rietveld 2014). The organism's actions aim at keeping its own free energy—the free energy associated with the self-model—to a minimum on average and over time. The actions that are inferred by the self-model are actions that contribute in the long run to keeping its own free energy to a minimum—that is to ensuring the organism remains in the bodily states it expects to occupy given its way of life in the niche it inhabits.

Through active inference—that is to say through cycles of perception and action in which the organism engages with the relevant affordances of its environment—the organism actively and continuously produces a distinction between itself and its environment. In the enactive theory, living systems are characterized in terms of autonomy (Varela 1979; Thompson 2007; Moreno and Mossio 2015; Di Paolo et al. 2017). An autonomous system is defined as a system that has the properties of (1) operational closure and (2) precariousness. Operational closure is a property of networks made up of processes that stand in mutually enabling relations. This has the consequence that the self-production of the network as a concrete unity and its distinction from the surrounding environment is due to the mutually enabling relations that hold between the processes of which the network is composed. The component processes are mutually enabling in the sense that the individual processes depend on each other recursively for their ongoing generation and re-generation (Moreno and Mossio 2015). The operational closure of the network is "precarious" in the sense that the activity of each individual constituent process is sustained through its relations to other constituent processes. Considered as an isolated process it will tend to run down and compromise the organization of the network as a whole.

According to the free energy principle, the autonomy of living systems is a consequence of the inferential processes of free energy minimisation (Kirchhoff 2017; Allen and Friston 2017; Kirchhoff and Froese 2017). Through active inference the organism establishes and sustains a boundary [technically referred to as a Markov blanket[13] (Friston 2013;

[12] For an extended treatment of the role of attention in free energy minimisation see Feldman and Friston (2010). Briefly, attention in this framework is understood in terms of precision–estimation where precision relates to the reliability of a prediction error signal. When a prediction error signal is assigned a low probability because the brain has low confidence in the signal, top-down predictions are allowed to influence future processing and control what the organism does next. Conversely, when prediction error signals are estimated to be reliable and are assigned a high-probability, further processing will be influenced by this prediction error either through acting so as to generate bodily states that match with predictions, or by updating top-down predictions of the self-model. Attention is thus assigned the role of modulating the relative influence of prediction error signals in relation to top-down predictions based on estimations of uncertainty. Attention will undoubtedly prove to be an important part of the account of consciousness and mineness I am developing in this paper. It plays a crucial role in minimisation of *expected* free energy which will be shown in Sect. 4 to be associated with consciousness. Unfortunately, exploring the relation between attention and mineness is beyond the scope of what I can do in this paper. On the role of precision–estimation in processes of expected free energy minimisation see Kiverstein et al. (2017).

[13] The concept of the Markov blanket is borrowed from causal graph theory where it refers to a set of nodes that make up a network such that for a given node X, the behavior of X can be fully predicted by knowing the states of the other nodes that make up the network. The

Hohwy 2016; Kirchhoff 2017)] separating their own internal dynamics from their external surroundings. The boundary is brought about through the circular causal interactions of perception and action as the organism engages with the relevant affordances of its environment. It is only if the boundary is sustained over time that the organism will avoid its own dissipation. This maintaining of the boundary is a consequence of the organism maximising the evidence for a self-model, or equivalently minimising its own free energy. Its cycles of perception and action are as we have been explaining, driven by this imperative. Allen and Friston express this idea eloquently:

> Simply put, an organism persists in virtue of having internal states which cause surprise-minimising, evidence maximising actions; these in turn maintain the partitions (of internal dynamics from the external surroundings) which is a necessary condition for existence. (Allen and Friston 2017, p. 16, parentheses are my addition).

Thus, the free energy principle tells us how living systems might sustain their own operational closure under precarious conditions in their dynamic coupling with the environment. Organisms do so by acting so as to keep the free energy of the self-model they instantiate to a minimum. Given a dynamic environment in which all manner of unpredictable occurrences take place there is never any guarantee of success. The organism must continuously adjust its internal dynamics to the random fluctuations of its environment. This can be done in two ways, either by acting so as to fulfil the organism's predictions or by updating the self-model so as to change what the organism expects so as to better reflect the dynamics of the organism's coupling to the environment. Either way the organism regulates its coupling to the environment so as to bring about the conditions of its own existence. The organism is the author of its own bodily states; it "actively brings about the conditions of its own survival" (Allen and Friston 2017, p. 16). In acting so as to minimise the discrepancy between the predictions of the self-model and the events that ensue, it is maximising the evidence for the self-model it has become. Since what is being modelled is the organism's own existence, what this amounts to is the organism acting so as to maximise the probability of its own existence.

Any system that has autonomy will also qualify as an agent that has its own individual point of view upon the world (Di Paolo et al. 2017). Relative to this point of view the environment has affective significance in terms of how it bears on the organism's self-produced identity. Organisms enact values, purposes and norms which are of their own making in the sense that they originate in processes of self-individuation (i.e. free energy minimisation) to which the organism owes its existence. The organism in its practical engagement with the environment makes distinctions between good and bad, better or worse, risk or opportunity (Di Paolo 2005; Thompson 2007; Rietveld 2008; Kiverstein et al. 2017). Perception and action are thus laden with affect: they possess an affective significance that stems from the organism's need to preserve its own operational closure under precarious conditions. Perception includes evaluations of how the organism is fairing in the world and where the opportunities and threats are to be found. An organism's affective states attune them to what is salient.

We can thus think of active inference in terms of readiness for action (Kiverstein 2017). What the agent is ready to act on are relevant affordances. The relevance of an affordance is determined by the organism's need to preserve its own operational closure. It should be noted that the concept of operational closure generalises well beyond homeostasis—it doesn't only relate to the internal physiological conditions of the body such as glucose and oxygen levels in the blood, breathing, body temperature, heart rate, and so on. Autonomy operates at multiple, interacting levels of organisation as organisms grow in complexity. As the organisation of the autonomous system becomes less bound to its immediate metabolic needs, so the possible meaningful relations the organism can stand in to the environment becomes less tightly bound to the here and now. The organism becomes sensitive to tendencies and trajectories that constitute the dynamical configurations of the organism–environment system, and their consequences for its precarious existence (Di Paolo et al. 2017, pp. 228–229). It becomes sensitive to the tendencies and trajectories in the evolution of its own states that stretch steadily further through time. The nervous system can then be thought of as generating and sustaining stable and recurrent patterns of sensorimotor engagement with the environment. These patterns of engagement with the environment exhibit just the same properties of operational closure and precariousness as we find in the more basic processes of homeostasis. The argument of this section is that patterns of sensorimotor engagement owe their operational closure to processes of free energy minimisation.

Given the connection between autonomy and free energy minimization I have been drawing out in this section, it follows that the organism's norms and purposes have their roots in active inference. In active inference, it is the actions that will contribute to the sustaining of the organism's

Footnote 13 (continued)

states of neighbouring nodes fix the state of the target node in a way that is not conditioned by (i.e. is conditionally independent from) all other states of the system. The nodes of the network in this sense constitute a Markov blanket around them that shields them from the activity in the rest of the system (Pearl 1988).

operational closure in its coupling to a dynamically changing environment that are inferred. We can see this by briefly considering how free energy as a mathematical quantity is kept to a minimum through active inference. Free energy is a function of two probability distributions. The "generative density" is a probability distribution that specifies the joint probability of sensory states S and states of the environment E conditioned on a self-model M. It maps the statistical relations that hold between the bodily states the organism expects itself to be in as a consequence of acting on the affordances of its econiche. The recognitional or variational density is typically explained as "encoding" the organism's approximate posterior beliefs about the hidden environmental causes of its bodily states. I've argued above however that in active inference the organism is not so much inferring the hidden external causes of its observations but is instead inferring the causal dynamics of agent–environment interactions. Thus, recasting this talk of the variational density as "encoding beliefs" in ecological and enactive terms, I suggest we think of the recognitional or variational density as the states of action readiness that relate to relevant affordances. What is inferred is thus the expected surprise associated with bodily states that arise as a consequence of responding to relevant affordances.[14] The quantity of free energy is determined by the generative density, the recognitional density and the current sensory states. Through active inference, each of these quantities changes in such a way to minimise free energy expected under a course of action.[15]

The central idea of the ecological enactive interpretation of active inference is that this process of active inference doesn't depend only on what is happening inside of the brain. It is instead to be understood at the level of the global dynamics of the whole organism as the organism regulates its coupling to the environment in such a way as to sustain its own operational closure across multiple levels of organisation.

We are finally in a position to assess how mineness might be understood in terms of active inference. I suggested above in Sect. 1 that mineness emerges in systems that combine their own purposive agency with processes of sensorimotor integration. We saw how cycles of perception and action (which I've just been explaining in terms of processes of active inference) are self-specifying. They are self-specifying because of the systematic relation between sensing and moving realised through the perception–action cycle. As the organism moves so its movements have consequences for what it senses, and what it senses has consequences for how the organism moves. Sensing and moving are in this way co-determining. By systematically tracking the relation of what it senses (reafferent feedback) to how it is moving (efference) the organism's perception comes to be self-specifying. Its perception comes to be anchored to a particular body, the organism's own body, and thus to a particular individual perspective on the world.

I've argued however that self-specifying perceptual states are not sufficient for mineness. In addition what is needed is that the movements of the organism be "purposive": they must be under the organism's control, regulated on the basis of the organism's own internal bodily conditions. Interoception is responsible for the organism's sense of the physiological condition of its own body from the inside such as hunger or fatigue. It provides the organism with a "feeling from within" of its own body as the perspective from which it practically engages with the world. As we've just seen, it is relative to this perspective that engagement with the environment can be going well or badly, can be risky or provide opportunities for fulfilment. I've been arguing that it is the sustaining of operational closure at multiple levels of autonomy through processes of free-energy minimisation that binds and integrates interoception together with processes of perception and action.

Consider as an example the alertness and concentration I need to sustain as I write this paper. We can think of this in terms of the self-model generating a top-down prediction that meets up with bottom-up sensory inputs that reflect my current state of fatigue and tiredness. My body may then take action to restore my desired level of alertness by predicting that I have a mug of coffee in my hand since this is how I typically deal with lack of concentration at this time of day. Motor control processes will then initiate the actions of making coffee that fulfil the prediction that I currently have a cup of coffee in my hand. If all goes well the prediction that I am alert and focussed will turn out to be self-fulfilling. Notice however that this is a prediction that doesn't only concern the inner states of my body. It is a prediction that ties together my internal physiological condition in the form of my level of fatigue with my perception and action in such a way as to minimise overall prediction error across the sensory hierarchy.

The ecological enactive interpretation of active inference I have developed in this section thus shows how processes of sensorimotor integration of the type that delivers self-specifying perception and purposive agency could be combined in a single cognitive package. This single cognitive package is a natural consequence of active inference as the means by

[14] My thanks again to an anonymous reviewer for this suggested formulation.

[15] This necessarily entails a prediction of the future as discussed in Sect. 3. Interestingly, it turns out that expected free energy can always be expressed in terms of epistemic and pragmatic affordances. The epistemic part resolves uncertainty about the niche–agent interaction; while the pragmatic part rests upon the prior beliefs of the agent that constitute its self-model. I thank an anonymous reviewer for this important observation. For further discussion see Friston et al. (2015).

which the organism sustains its own operational closure in dynamically coupling to its eco-niche. Can we end the paper here then and conclude that active inference is what delivers mineness? Can we say that any living system must have a basic form of self-awareness because any living system must sustain its own operational closure through free energy minimisation? Unfortunately, such a conclusion would seem premature. All living systems as we began by noting must engage in free energy minimisation. This much follows from the free energy principle. They will do so through processes of active inference. But not all living systems are conscious. Thus active inference cannot be sufficient for mineness. Something extra must still be needed.

5 Inferring Ourselves?

Karl Friston has recently argued that any conscious creature must be able to make inferences about the consequences of its own actions in the future (Friston 2017, 2018). It must be capable of making what he describes as "temporally thick" inferences. "Temporal thickness" refers to the capacity to make inferences not only about the present but also about the past and the future. In order to be conscious Friston suggests an organism must be able to project itself through time—it must be able to select actions that minimise its expected or future uncertainty. It is the prospective dimension of active inference that he takes to be crucial for consciousness. Such a living system that is able to engage in "proactive, purposeful inference about its own future" based on a temporally thick self-model will be able to infer actions that will minimise the surprise expected as a consequence of its actions (Friston 2017, 2018).[16] Consciousness may be associated with processes of active inference that selects actions which minimise non-actual but possible future surprise. A conscious creature can infer actions based on its own future prospects. Loss of consciousness is due to loss of the temporal thickness of a self-model such as can be found in a deep coma (Friston 2017). Interestingly one of the tests for a minimal consciousness is the ability to engage in imaginary actions such as playing tennis (Owen et al. 2006; Shea and Bayne 2010). On Friston's proposal this would make good sense as a test for minimal consciousness. The ability to engage in this kind of imagination is a sign of the re-emergence of a temporally thick self-model.

Friston's hypothesis that temporal thickness may be what needs to be added to active inference to yield mineness makes good sense in the light of what was claimed earlier about autonomy operating at multiple levels of organisation. There it was argued that as organisms grew in the complexity of their biological organisation, so the possible meaningful relations they could take up to the environment likewise expanded beyond the here and now. Minimal agents such as bacteria regulate their coupling to the environment on the basis of metabolic relevance. They move away from potentially noxious substances and towards concentrations of nutrition but are confined to acting more or less in the here and now. But more complex agents are not only sensitive to the conditions of the environment here and now. They are able to sustain recurrent patterns of interaction with the environment and are sensitive to the possible evolution of the agent–environment system and the consequences of this evolution for their own precarious existence. Temporally thick inference thus emerges naturally out of growth in the complexity of an organism's existence. Friston's hypothesis is thus very much in keeping with the claim I made in Sect. 1 that mineness emerged relative early in evolutionary history because of a connection between action control and subjectivity.

Along convergent lines to Friston, Gallagher (2017) has recently argued enactive perception must have a temporal thickness. As my hand move towards the cup of coffee I am reaching to grasp, my arm goes through a sequence of different postures. At each moment my movement is unfolding along a trajectory because of the cup I am aiming to reach. There is thus a retaining in perceptual presence of the cup's affordances—its possibilities for action—to which my movements are coordinating and adjusting. At the same time my movements are unfolding in a way that anticipates my taking hold of the cup of coffee to drink from it. My movements thus unfold along a particular trajectory based both on a retention of my body's configuration in relation to the environment, and an anticipation of where my movement is heading next. Similarly, perception is not a "knife-edge impression of the present." Perception instead arises with what Gallagher describes as an "empty anticipation" that is either fulfilled or not fulfilled. This empty anticipation is in turn constrained by what Gallagher calls "retention" of what was just anticipated (following Husserl).[17] Temporality

[16] The term "surprise" is being used here in the information-theoretic sense as outlined above to refer to bodily states that have a low probability of occurring on average and over time. Surprising states are unpredictable and they threaten the organism's integrity should the organism find itself in them.

[17] In this paper Gallagher is developing an enactive interpretation of Husserl's genetic analysis of the phenomenological structure of internal time consciousness. In Husserl's analysis every conscious experience has a threefold temporal structure (Husserl 1991). It comprises a retentional part that presents the subject with what has just past, a primal impression that is constantly arising anew in the now, and a protentional or anticipatory element that is directed at what is about to happen in the near future. Husserl's phenomenological analysis of time consciousness has been shown by Dan Zahavi to account for how every conscious experience can include a dimension of mineness (Zahavi 2005). This is because what is retained is an entire phase of experience that has just past with its own threefold temporal structure. Retention thus makes it the case that every experience is always

is what explains the directedness of both consciousness and action towards something in the environment. Consciousness as enactive is to be understood as an "I can" that is as an "apprehension of the possibilities or affordances in the present." (Gallagher 2017, p. 13) There would be no engagement with affordances were perception to only present an animal with a knife-edge present. To apprehend and be sensitive to possibilities, a perceiving animal needs prospection—it needs to have experiences that reach out into the future anticipating what could be. This is just what it takes to perceive possibilities. Gallagher does not spell out whether perception of possibilities would be possible without retention. However since what is retained is just the fulfilled or unfulfilled protention that has just past, we can infer that it would not. Perception without retention would be perception that is unconstrained by what was previously anticipated. But we have just argued that there can be no perception of possibilities without prospection.

Friston's proposal to explain consciousness in terms of temporally thick self-models fits well the argument of this paper. Can we conclude then that the self that is intrinsic to conscious experience is the product of inferential processes? Do I infer my existence as a self? There is an important sense in which such a conclusion is indeed implied by the theory currently under consideration. I've argued that living systems embody a probability distribution conditioned on a model that identifies the bodily states the organism regularly occupies, and indeed must occupy if it is to continue to exist. This probability distribution is thus conditioned on a model that is inferred as the best explanation of the organism's own existence in its eco-niche. But if this characterisation of life as an inferential process is correct, it follows that selves must likewise be the outcome of inferential processes. What Friston adds to the ecological and enactive interpretation of active inference I've been developing is the requirement that the self-model have temporal thickness. But we've seen that this requirement falls naturally out of the growing complexity of organisms as they acquire the capacity to sustain operational closure over longer time scales. The temporal thickness requirement seems to follow straightforwardly from the ecological and enactive account I have proposed.

However, one might worry that the resulting account of mineness as a conclusion of probabilistic inference implies the possibility of making all kinds of mistakes about ourselves. Isn't such a conclusion at least *prima facie* in tension with thinking of mineness as intrinsic to the structure of conscious experience of the world? In the final section I will consider and respond to this objection.

6 Are Experiences Logically Immune to Error Through Misidentification?

The question of whether I am the subject of a given experience ordinarily doesn't arise. My experiences have what some philosophers have called the logical feature of immunity to error through misidentification (Shoemaker 1968; Gallagher 2000). They have this feature because of the peculiar experiential access I have to my own experience. I can access my experiences from the inside immediately and directly without the need for self-identification. I can do so because my experiences have what I've been calling "mineness" built into them intrinsically.

To see this further, contrast the following two scenarios. In the first I observe my body in the mirror. I am the *object* of a conscious experience. In order to recognise myself I have to identify myself with the person seen in the mirror. There is however the logical possibility of making a mistake, or even of failing to recognise oneself, as in Mach's famous story of the shabby pedagogue. Mach recounts the story of climbing onto a tram and thinking to himself about someone he saw at the other end of the tram: "That man is a shabby pedagogue" (Mach 1914). Later he realised that the man he was looking at was in fact a reflection of himself in a mirror at the back of the tram. Mach failed to recognise himself when seeing his image reflected in the mirror.

In the second scenario, I am the *subject* of a conscious experience. I am currently in pain. There is no question about whether the pain belongs to me or someone else. This is not something about which I can be mistaken. That I am the subject of this experience is settled by the very phenomenological structure of the experiences. My pain experience is an experience *for me* first and foremost, and for others only secondarily.

Does the view of the self that follows from the free energy perspective require us to revise the core claim of the phenomenological theory that experiences are logically immune to error through misidentification?

As already suggested at the end of the last section, it might be thought one must answer this question affirmatively. Consider the rubber hand illusion for instance. The brain treats the rubber hand as part of the subject's body as result of processes that infer the model of the body that best accounts for current multisensory stimulation. This results in me misidentifying something as being mine (the rubber hand) that is not mine. Experiments like the rubber hand illusion seem to show that our experience of what is mine and what is not mine is highly flexible and malleable. We are all too willing it seems to succumb to "weird metamorphoses

Footnote 17 (continued)

itself experienced. This intriguing convergence of ideas is something I hope to return to in future work.

of our customary ways of experiencing our bodies" (Hohwy 2013, p. 236).

In the rubber hand illusion, surprise is evoked by simultaneously feeling one's hand being touched and observing touch to an external object. The hand the subjects sees receiving tactile stimulation is different from the hand she feels being touched. The hand she sees receiving tactile stimulation is the rubber hand. The hand she feels is her own hand. There are thus two possible hypotheses that might possibly make sense of this incongruent sensory information. The first correctly hypotheses that vision and touch are taking place on different hands at different locations. My own hand is a part of me, the rubber hand is not. The second incorrectly hypotheses that vision and touch are taking place at the same location in the same hand. To explain away the prediction error arising from the incongruent multisensory information the brain must infer a hypothesis about the current features of the self that minimise overall surprise. The conclusion the brain reaches is that vision and touch are taking place at the same location in the same hand. This hypothesis best fits one's current sensory information and one's prior beliefs because it does the best job of resolving the conflicting sensory information relative to other of the subject's beliefs. The subject believes for instance that multisensory inputs tend to occur in one and the same object. Giving up on this belief would generate more prediction error in the long run. The belief that the rubber hand is not a part of my body is thus ignored. It is weighted as less probable than a hypothesis that takes the rubber hand to be a part of me (Hohwy and Paton 2010; Hohwy 2013, Apps and Tsakiris 2014; Limanowski and Blankenburg 2013).

Jakob Hohwy has argued that in active inference the brain models its causal interventions in the world based on the assumption of a single, coherent bodily trajectory that gives rise to a single coherent flow of multisensory input (Hohwy 2013, chap. 10 and 11; also see; Michael and Hohwy 2018). The brain adopts this assumption because it cannot use two or more conflicting hypotheses as its basis for sampling the world. It thus needs to select a single hypothesis that has the highest overall posterior probability about the state of the body in relation to the world in active inference. Sometimes this requires the brain to opt for a hypothesis about the self that is blatantly false. The brain prefers such a false hypothesis because it does the best overall job of combining what is already believed with current evidence so as to reduce overall prediction error. This is the case as we have seen in the rubber hand illusion. According to Hohwy the self is among the hidden causes of sensory input the brain must represent as a part of the process of inferring a model that does the best job of explaining away prediction errors.

Katerina Fotopoulou has defended a similar proposal in a series of recent papers (Ciaunica and Fotopoulou 2017; Fotopoulou and Tsakiris 2017). She has argued for an explanation of such representations of the self as the outcome of a process she calls "embodied mentalisation"—"the inferential brain process by which primary sensorimotor and multisensory signals are progressively integrated and schematised to form multiple, predictive models of our embodied states in given environments" (Fotopoulou and Tsakiris 2017, p. 8). For example, the sensory flows of information that originate in one's own body tend to stand in relations of temporal synchrony. Thus when two or more sensations belonging to different sense modalities (say touch and vision as in the rubber hand illusion) are observed to stand in relations of temporal synchrony, the brain will infer that the source of these sensory signals is most likely the subject's own body. Fotopoulou refers to this process as "mentalisation" because the organisation and schematisation of bodily signals applies to any body, not only my body but also to the bodies of others (Ciaunica and Fotopoulou 2017, p. 6). The same inferential processes that lead to self-consciousness also lead to the understanding of others because the body that one infers as most likely one's own is just one embodied subject among others. In her work mineness is explained as emerging out of such inferential processes of organising and schematising multisensory bodily state. Some of these multisensory signals originate from inside of the body and some from physical contact with what is outside of us. The self is just "the most parsimonious and accurate explanation for sensory inputs" (Fotopoulou and Tsakiris 2017, p. 10).

Along similar lines to these authors, I've been arguing that mineness is rooted in processes of active inference in which the organism acts so as to minimise its own expected future surprise. Must I agree with them that the self is just the outcome of the brain's abductive inferences? Such a conclusion doesn't sit well with my phenomenological starting point in this paper. Far from showing that mineness is intrinsic to phenomenal consciousness, the idea of the self as an inferential construct seems to threaten and provide a challenge to the phenomenological theory. The phenomenological theory requires that experiences be logically immune to error through misidentification. Experiences do not require us to identify ourselves as their subjects because they have mineness intrinsically built into them. But the upshot of the inferentialist account of the self is that experiences lack this feature of immunity to error, or at best they have this feature only contingently, not logically (Gallagher 2012). We can make all kinds of mistakes about what is or is not mine or me, and we can do so precisely because mineness is the outcome of inference.

I will finish up by showing how the ecological and enactive interpretation of active inference I have been developing does not have this consequence. The self is not a hidden cause of sensory input that stands in need of inference in a way that makes trouble for the logical immunity claim.[18]

[18] I am extremely grateful to an anonymous reviewer for pressing me to say more on this point. Their probing questions helped me to refine the differences between my own ecological–enactive interpretation

Consider again the rubber hand illusion. When the subject makes a mistake about the position of her hand mistaking the position of her own hand for the position of the rubber hand she does so by switching perspectives on her body. Her perspective on her own hand is non-observational. Her own hand is the hand through which she can take hold of things and make use of them. The rubber hand is one she relates to from an observer's standpoint. It is something she can observe from a spectator's standpoint.

Recall the distinction with which I began this section between an experience in which I am the object of an experience, and an experience in which I occupy the subject position. In the case of my own arm, I have an experience of my body as subject. My arm is a part of my first-person perspective on the world. It is that through which I experience the world. In the second case, I have an experience of my body as object. I can make a mistake about whether a body or a part of my body that I see is my body just as we saw earlier in the example from Mach. This kind of mistake arises because identification of which body is my body seems to be required. However in the case of experiencing my body as subject no such identification is required. What proprioception and kinesthesis deliver is the sense I have of what I can do with my body in space. They form the basis for my body's readiness for action in relation to relevant affordances.

In the rubber hand illusion it is the *recognition* of what is a part of my body that is at stake. Recognition that targets my body does indeed result in awareness of my body, but it is an awareness of my body as an object among other object. This is more or less made explicit in the enfacement illusion in which one looks in a mirror and observes the face of a stranger being stroked in synchrony with one's own face (Tajadura-Jimenez et al. 2012). The enfacement illusion occurs when you misrecognise the reflection of someone else's face as your own face because of processes of multisensory integration. In both the rubber hand and enfacement illusions you have an illusory experience of your body as an object.

What is being investigated in these studies is, I am suggesting, an awareness of the body as object. As Shaun Gallagher has nicely shown immunity to error through misidentification may well fail for this type of awareness of the body without such a result implying that it fails more generally for the awareness we have of ourselves as subjects (Gallagher 2012). Such a conclusion would follow if awareness of the body as subject or mineness was also a hypothesis arrived at through a process of inference to the best explanation, the brain's best guess about the cause of its current sensory input. Does this follow from the argument that has been given?

I've argued that mineness is *intrinsic* to life in organisms of sufficient complexity. Life is indeed an inferential process—it is a probability distribution conditioned on a self-model. Should we conclude then that life is just a hypothesis constructed by an organism's brain—the brain's best guess about the hidden causes of its current sensory input?

I suggest this is the wrong conclusion to draw from the arguments of this paper. Recall that I have argued that a consequence of free energy minimisation is that the organism sustains its operational closure across multiple levels of autonomy. It thereby relates to a meaningful environment of action possibilities because the environment is made up of possibilities that bear either positively or negatively on its processes of self-individuation. The consequence of free energy minimisation is thus the sustaining of operational closure across multiple levels of biological organisation. It is the bringing forth or enactment of domains of affective significance in which distinctions can be made between possibilities that are good or bad for the organism.

Mineness is I've argued intrinsic to processes of free energy minimisation that sustain autonomy once these processes have reached the right level of complexity. More specifically, once an organism is capable of acting so as to minimise its own expected free energy, such a living process will now have developed its own lived first person perspective on its environment. An organism will then experience its environment through its body. Its body will serve as the origin of an egocentric spatial framework from which it relates to its surroundings. It will undergo intentional states that are directed at the world, and that are at one and the same time also *for itself*, in the sense of being self-conscious (Sartre 2003).

7 Conclusion

The key issue we've arrived at in this final section is whether the inferential processes that form the basis for life have to be understood in representational terms. It is commonly taken for granted, and sometimes explicitly argued that Bayesian inference is a process that is carried out on probabilistic representations.[19] Bayesian inference is understood as something that neural processes approximate by updating the prior beliefs the brain encodes in its patterns of neural

Footnote 18 (continued)

and other theorists that have also offered explanations of the self in terms of processes and free energy minimisation (e.g. Hohwy 2013; Apps and Tsakiris 2014; Fotopoulou and Tsakiris 2017).

[19] See e.g. Gladziejewski (2016) and Kiefer and Hohwy (2017) for detailed arguments to this effect.

activation based on new sensory evidence so to maximise the probability of a posterior hypothesis.

I have been arguing for an ecological and enactive interpretation of the free energy principle in this paper. On my interpretation it is the life of the organism in its eco-niche that is understood as an approximation of Bayesian inference, not only the processes in the organism's brain. The organism is a self-model that acts so that on average and over time it avoids surprise, engaging with the environment selectively so as to maximise the evidence for its self-model as a model of its own existence. A self-model maps statistical relationships that hold among bodily states, the organism's actions, and the affordances of its econiche. It maps the statistical structure of organism–environment interactions, not the hidden causes of sensory input conceived of independent of the agent. Minimising free energy is achieved through the process of active inference. A consequence of active inference is that the organism is able to sustain its own operational closure at multiple levels of biological organisation all the way up to the recurrent patterns of interaction and engagement with the affordances of the environment.

I've argued that processes of free energy minimisation of the right complexity may thus be sufficient for the emergence of mineness. Free-energy minimisation is however something the organism accomplishes through a process of active inference—an inferential process that unfolds within an organism–environment system. It follows that lifeforms that have mineness as a consequence of active inference likewise have to be understood in relation to the eco-niche they inhabit. The forms of operational closure they sustain through their activities relate to their way of life in their niche.

The human way of life is socio-cultural. People regulate their interactions with the environment more generally based on what they care about, and how the world matters to them (e.g. their friends and family, the work projects, rituals and social practices they take part in). Humans don't only expect bodily states that are consistent with maintaining their internal physiological milieu within a constrained range of values through homeostasis (Seth 2013). They expect to occupy bodily states that relate to their way of life more generally—the habits and practices they engage in regularly and repeatedly. In other words, human agents expect to occupy bodily states that relate to their own flourishing in the niche they construct in part through cultural and social practices. They allow themselves to be drawn into action by relevant affordances that sustain their values and what matters to them. The result of free-energy minimisation is that they further their own flourishing both as individuals, but also crucially in relation to others.

This paper is part of a special issue on the topic of the relational self. Thus it is appropriate to close with a brief comment on the implications of the arguments of this paper for a relational theory of the self. The bodily self has been argued to be a relational self because it is inferred on the basis of processes of embodied mentalisation that integrate bodily signals including signals that come from the outside from other bodies in close physical proximity (Ciaunica and Fotopoulou 2017; Fotopoulou and Tsakiris 2017). The processes of multisensory integration that lead to the development of bodily self-awareness in infants include the integration of signals that arise from interpersonal touch (e.g. hugging, breastfeeding, carrying etc.). The bodily signals that are organised and schematised in processes of embodied mentalisation thus need not originate exclusively from inside of the body, but will include signals that originate from the infant's affective bodily contact with its caregivers.

It might be objected that the account of embodied mentalisation proposed by Fotopoulou and her colleagues implies a neurocentric account of the self that is very much at odds with the arguments of this paper. Embodied mentalisation is after all a process of schematising multisensory information that occurs within the brain of an individual. However, if the arguments of this paper are along the right lines embodied mentalisation as a process of free energy minimisation is better interpreted in ecological–enactive terms. The interoceptive bodily signals that are organised and schematised are the bodily states that need to be kept within a tightly restricted range of values as part of the process of sustaining operational closure. The self that emerges out of processes of free energy minimisation is a relational self because self and other are co-dependent and co-determining. The upshot of Fotopoulou's important research is that the self that is implied by processes of free energy minimisation is not an isolated individual but a self that cares deeply about its interactions with others. The self that is sustained through free energy minimisation receives its nourishment and grows out of its interpersonal relationships.

Funding This research was funded by the European Research Council Starting Grant awarded to Prof. Erik Rietveld (Grant Number 679190).

Compliance with Ethical Standards

Conflict of interest No conflict of interest.

Research Involving Human and Animal Participants This article does not contain any studies with human participants or animals performed by any of the authors.

References

Allen M, Friston K (2017) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese. https://doi.org/10.1007/s11229-016-1288-5

Anderson M (2014) After phrenology: neural reuse and the interactive brain. MIT Press, Cambridge

Apps M, Tsakiris M (2014) The free energy self: a predictive coding account of self-recognition. Neurosci Biobehav Rev 41:85–97

Bruineberg J, Rietveld E (2014) Self-organisation, free energy minimisation, and optimal grip on a field of affordances. Front Hum Neurosci 8:599

Bruineberg J, Kiverstein J, Rietveld E (2017) The anticipating brain is not a scientist. The free energy principle from an ecological-enactive perspective. Synthese. https://doi.org/10.1007/s11229-016-1239-1

Calvo P, Friston K (2017) Predicting green: really radical (plant) predictive processing. J R Soc Interface 14(131):1742–5662

Campbell JO (2016) Universal darwinism as a process of Bayesian inference. Front Hum Neurosci 10:49

Christoff K, Cosmelli D, Legrand D, Thompson E (2011) Specifying the self for cognitive neuroscience. Trends Cogn Sci 15(3):104–112

Ciaunica A, Fotopoulou K (2017) The touched self. Psychological and philosophical perspectives on proximal intersubjectivity and the self. In: Durt C, Fuchs T, Tewes C (eds) Embodiment, enaction and culture: investigating the constitution of the shared world. MIT Press, Cambridge

Clark A (2003) Natural born cyborgs: minds, technologies and the future of human intelligence. Oxford University Press, Oxford

Dewey J (1896) The reflex arc concept in psychology. Psychol Rev 3:357–370

Di Paolo E (2005) Autopoiesis, adaptivity, teleology and agency. Phenomenol Cogn Sci 4:429–452

Di Paolo E, Buhrmann T, Barandiaran XE (2017) Sensorimotor life: an enactive proposal. Oxford University Press, Oxford

Feldman H, Friston K (2010) Attention, uncertainty and free energy. Front Hum Neurosci 2(4):215

Fotopoulou K, Tsakiris M (2017) Mentalising homeostasis: the social origins of interoceptive inference. Neuropsychoanalysis 19:71–76

Friston K (2010) The free energy principle: a unified brain theory? Nat Rev Neurosci 11(2):127–138

Friston K (2012) A free energy principle for biological systems. Entropy 14(11):2100–2121

Friston K (2013) Life as we know it. J R Soc Interface 10:20130475

Friston K (2017). The mathematics of mind time. Aeon. https://aeon.co/essays/consciousness-is-not-a-thing-but-a-process-of-inference

Friston K (2018) Am I self-conscious. Front Theor Philos Psychol. (**Manuscript under review**)

Friston K, Stephan K (2007) Free energy and the brain. Synthese 159(3):417–458

Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G (2015) Active inference and epistemic value. Cogn Neurosci. https://doi.org/10.1080/17588928.2015.1020053

Fulda FC (2017) Natural agency: the case of bacterial cognition. J Am Philos Assoc 3:1–22

Gallagher S (2000) Philosophical conceptions of the self: implications for cognitive science. Trends Cogn Sci 4(1):14–21

Gallagher S (2005) How the body shapes the mind. Oxford University Press, Oxford

Gallagher S (2012) First-person perspective and immunity to error through misidentification. In: Miguens, Preyer (eds) Consciousness and subjectivity. Ontos Verlag, Frankfurt, pp 187–214

Gallagher S (2017) The past, present and future of time consciousness: from husserl to varela and beyond. Constr Found 13(1): 91–97.

Gibson JJ (1979) The ecological approach to visual perception. Houghton Lifflin, Boston

Gladziejewski P (2016) Predictive coding and representationalism. Synthese 193(2):559–582

Godfrey-Smith P (2016) Other minds: the octopus and the evolution of intelligent life. Harper Collins, London

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

Hohwy J (2016) The self-evidencing brain. Noûs 50(2):259–285

Hohwy, J. (2017). How to entrain your evil demon. In: Metzinger T, Wiese W (eds) Philosophy and predictive Processing. MIND Group, Frankfurt am Main. https://doi.org/10.15502/9783958573048

Hohwy J, Paton B (2010) Explaining away the body: experiences of supernaturally caused touch and touch on non-hand objects within the rubber hand illusion. PLoS ONE 5(2):e9416

Husserl E (1991) *On the phenomenology of the consciousness of internal time* (1893–1917). (trans: Brough J). Kluwer Academic, Dordrecht

Kiefer A, Hohwy J (2017) Content and misrepresentation in hierarchical generative models. Synthese. https://doi.org/10.1007/s11229-017-1435-7

Kirchhoff M (2017) Autopoiesis, free energy and the life-mind continuity thesis. Synthese. https://doi.org/10.1007/s11229-016-1100-6

Kirchhoff M, Froese T (2017) Where there is life there is mind: in support of a strong life-mind continuity thesis. Entropy 19(169):1–18

Kiverstein J, Miller M, Rietveld E (2017) The feeling of grip: novelty, error dynamics and the predictive brain. Synthese. https://doi.org/10.3389/fnhum.2015.00237

Klein C (2015) What the body commands: an imperative theory of pain. MIT Press, Cambridge

Legrand D (2006) The bodily self: the sensorimotor roots of pre-reflexive self-consciousness. Phenomenol Cogn Sci 5:89–118

Legrand D (2012) Self-consciousness and world-consciousness. In: Zahavi D (ed) The Oxford handbook of contemporary phenomenology. Oxford University Press, Oxford

Limanowski J, Blankenburg F (2013) Minimal self-models and the free energy principle. Front Hum Neurosci 7:547

Mach E (1914) The analysis of sensations and the relation of the physical to the psychical. Open Court, Chicago

Metzinger T (2003) Being no-one: the self-model theory of subjectivity. MIT Press, Cambridge

Metzinger T (2009) The ego tunnel: the science of the mind and the myth of the self. Basic Books, New York

Michael J, Hohwy J (2018) Why should any body have a self? In: De Vignemont F, Alsmith A (eds) The subject's matter: self-consciousness and the body. Oxford University Press, Oxford

Moreno A, Mossio M (2015) Biological autonomy: a philosophical and theoretical enquiry. Springer, Berlin

Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD (2006) Detecting awareness in the vegetative state. Science 315:1221

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco

Powers W (1973) Behaviour: the control of perception. Aldine, Chicago

Ramstead MJD, Badcock PB, Friston KJ (2017) Answering Schrodinger's question: a free energy formulation. Phys Life Rev 24:1–16

Rietveld E (2008) Situated normativity: the normative aspect of embodied cognition in unreflective action. Mind 117(468):973–1001

Sartre J-P (2003) Being and nothingness: an essay in phenomenological ontology (trans: Barnes HE). Routledge, London

Seth A (2013) Interoceptive inference, emotion and the embodied self. Trends Cogn Sci 17(11):565–573

Shea N, Bayne T (2010) The vegetative state and the science of consciousness. Br J Philos Sci 61:459–484

Shoemaker S (1968) Self-reference and self-awareness. J Philos 65:555–567

Tajadura-Jiménez A, Longo MR, Coleman R, Tsakiris M (2012) The person in the mirror: using the enfacement illusion to investigate the experiential structure of self-identification. Conscious Cogn 21(4):1725–1738

Thompson E (2007) Mind in life: biology, phenomenology and the sciences of mind. Harvard University Press, Cambridge

Thompson E (2015) Waking, dreaming, being: self and consciousness in neuroscience, meditation and philosophy. Columbia University Press, New York

Tribus M (1961) Thermodynamics and thermostatics: an introduction to energy, information and states of matter with engineering applications. D. Van Nostrand, New York

Varela FJ (1979) Principles of biological autonomy. North Holland, New York

Zahavi D (2005) Subjectivity and selfhood: investigating the first-person perspective. MIT Press, Cambridge

Zahavi D (2014) Self and other: exploring subjectivity, empathy and shame. Oxford University Press, Oxford

Zahavi D (2017) Thin, thinner, thinnest: defining the minimal self. In: Durt C, Fuchs T, Tewes C (eds) Embodiment, enaction and culture: investigating the constitution of the shared world. MIT Press, Cambridge