Conscious self-evidencing

Jakob Hohwy
Cognition & Philosophy Lab
Centre for Consciousness and Contemplative Studies
Monash University, Melbourne, Australia

(Accepted for publication in *Review of Psychology & Philosophy*. Special issue on consciousness and predictive processing; Ed. by M. Miller, T. Schlicht, A. Clark.)

Abstract
Self-evidencing describes the purported predictive processing of all self-organising systems, whether conscious or not. Self-evidencing in itself is therefore not sufficient for consciousness. Different systems may however be capable of self-evidencing in different, specific and distinct ways. Some of these ways of self-evidencing can be matched up with, and explain, several properties of consciousness. This carves out a distinction in nature between those systems that are conscious, as described by these properties, and those that are not. This approach throws new light on phenomenology, and suggests that some self-evidencing may be characteristic of consciousness.

Keywords: self-evidencing; predictive processing; active inference; consciousness; self; perception

1. Self-evidencing is not sufficient for consciousness

A core notion of predictive processing is that existing, self-organising systems are self-evidencing (Hohwy 2016). Their internal states and their actions upon the environment are in the service of maximising the evidence for their own existence, for themselves. Equivalently, self-evidencing systems reduce uncertainty about their sensory input, given their model of the states they expect to occupy. This model defines what kind of system they are – in some sense *who* they are.

Self-evidencing *per se* is not an evolved function of some creatures. Rather, self-evidencing falls out of consideration of what it means for a self-organising system to exist. Specifically, it relates to a statistical understanding of existence of such (near equilibrium steady state) system in terms of the probability of occupying some and not other states. The free energy principle (Friston 2010, Friston 2019) argues that in a changeable world, active systems exist in virtue of managing to occupy their expected states; equivalently, they organise themselves and act in their environments to avoid surprising states, given their model (there is debate about the scope of the free energy principle; here I focus on self-organising, active systems). The free energy principle rests on a compelling argument that in order to do so, such systems focus on a quantity that is available for internal assessment, namely the free energy, which is a measure of the surprise of their sensory input given their model (often usefully simplified in the predictive processing framework as the long-term average of prediction error) (it is thus a transcendental argument that underlies the free energy principle, concerning the necessary conditions for the possibility of existence of self-organising systems, see (Hohwy 2020)). If the system is able to minimise free energy, then

they will implicitly minimise the surprise about their input, given what kind of model of the world they embody; minimising free energy is then equivalent to maximising evidence for the system's model of the world – self-evidencing. The appeal to free energy is critical because systems cannot directly assess the surprise of a given sensory input; this is because doing so directly would impossibly require the system knowing *a priori* which states it is supposed to occupy among all the indefinite number of possible states they could occupy.

The idea is then that self-organising systems essentially self-evidence when they, in a broad sense, perceive, change their internal states, and act upon the world. The reasoning from existence to free energy minimisation is conceptual (what is existence of self-organising systems?) and mathematical (proving that free energy bounds surprise, and thus model evidence). It is an austere vision of existence, which systems may or may not conform to, and self-evidencing does not convey a law of nature in the sense of quantifying over some domain (for discussion, see Hohwy 2020).

Assume that some self-organising systems in the world are conscious and some are not conscious. If indeed all self-organising systems are self-evidencing in the basic way described, then self-evidencing describes both conscious and non-conscious systems. Self-evidencing is then necessary for consciousness in self-organising systems in some fairly unspecific sense, but the fact that a given system is self-evidencing (and predictive processing) is in itself not sufficient for ensuring that the system is conscious. This presents a challenge for attempts to build a theory of consciousness directly on self-evidencing and the free energy principle. It is necessary to find a distinction within self-evidencing that can provide a demarcation between conscious and non-conscious systems. There are now several intriguing approaches to consider consciousness in terms of predictive processing and kindred ideas (Carhart-Harris, Leech et al. 2014, Rudrauf, Bennequin et al. 2017, Friston 2018, Williford, Bennequin et al. 2018, Clark 2019, Chang, Biehl et al. 2020, Friston, Wiese et al. 2020). Conversely, there are proposals for how to re-cast existing theories in terms of predictive processing such as heterophenomenology (Dołęga and Dewhurst 2020), global neuronal workspace theory (Hohwy 2013, Whyte 2019, Whyte and Smith 2020), versions of higher-order thought theory and metacognitive theories (Hohwy 2015, Sandved Smith, Hesp et al. 2020), attention-based theories (Marchi and Hohwy 2020), and integrated information theory (see Kanai, Chang et al. 2019). The assumption that some predictive processing systems can be non-conscious, and questions about demarcation, are rarely addressed head-on.

There is a different strategy available, which contrasts with building a new theory of consciousness. This strategy considers predictive processing broadly conceived the explanatory framework for consciousness science. It can be used to interpret the neural mechanisms revealed in the search for the neural correlates of consciousness, thereby throwing new light on different aspects of consciousness and properties of conscious phenomenology (Hohwy 2013: Ch 10, Hohwy and Seth 2020). There are some recent proposals that also move in this direction (Ramstead, Wiese et al. 2020, Wiese 2020, Wiese and Friston 2020, Ramstead, Hesp et al. 2021).

On this approach, the fact that self-evidencing in itself is insufficient for consciousness does not imply that self-evidencing is irrelevant for consciousness. Though the basic notion of

self-evidencing is austere, there are several distinct ways that different types of systems can minimise their free energy, or act to occupy unsurprising states. Some of these ways may be explanatorily relevant for consciousness, whereas others are not, allowing a demarcation in nature. That is, some specific processes for self-evidencing may be explanatorily relevant for specific properties of consciousness, while respecting the assumption that some systems are not conscious.

This strategy puts a list of phenomenological properties of consciousness first, focusing on properties worth caring about and worth explaining. The initial list may be incomplete, failing to capture all and only conscious systems. Eventually, the project of matching properties of consciousness to self-evidencing processes could be made complete, and then one could erect a theory of consciousness from self-evidencing. That more ambitious project is not undertaken here. Instead, after situating the project within consciousness science (Section 2), an initial list of properties of consciousness is identified, which it would be worth explaining (Section 3). Then diverse ways of self-evidencing are described, and matched up with the listed properties of consciousness (Section 4). Finally, consideration of what can be accomplished by this strategy (Section 5).

The outcome is not a comprehensive theory of consciousness. But it paints an interesting picture of core properties of consciousness: consciousness is formed through the way we internally model the external world, it unfolds in causal interaction with the world, but it is forged in deeply detached, internal processes driven by self-evidencing.

2. Self-evidencing in the science and metaphysics of consciousness

In consciousness science, the notion of consciousness is used in different ways, referring variously to awareness, awakeness, or sentience, or to either conscious state or conscious contents, or to what philosophers call phenomenal or qualitative states, or qualia, or to being conscious or what one is conscious of, or, to self-consciousness, self-awareness, or just the self. The varied use of the notion of consciousness is not immediately damaging to the scientific prospects of consciousness science. One of the often overlooked successes of consciousness science is the work done to parse out different senses of the term, making it easier to speak more precisely, and to stratify and operationalise research accordingly.

However, there remains a slippery, difficult-to-define notion of consciousness underlying most of the various uses of the term. Our inability to unequivocally define in functional or structural terms what this notion is lies at the heart of the philosophical debate about the metaphysics of consciousness. This includes arguments that the nature of consciousness must necessarily escape all scientific explanation, at least for sciences that operate in terms of the function and structure of natural processes. This is the difficult, 'hard' metaphysical problem of consciousness – the mind-body problem (Nagel 1974, Chalmers 1996, Levine 2001).

No-one, it is safe to say, has as yet solved the mind-problem problem, and it will not be attempted here. There are, rather, various ways to address or circumvent the metaphysical problem. One is to focus on what has been called the 'easy' problems of consciousness (metaphysically easy, yet still scientifically hard, see Chalmers 1995, Chalmers 1995): processes for integration of information, sensory discrimination, reportability, differences

between global states of consciousness and so on. These processes are not exclusive to consciousness but scaffold it in various ways. They are easier to explain because they can be given functional and structural definitions and explained in neuroscience or computational terms.

In spite of the murky centre of the notion of consciousness, there are also phenomenological properties that we take to be characteristic of conscious systems, and not of non-conscious systems. And these properties can be given functional or structural definitions that poise them sufficiently to be explanatory targets for naturalistic science. Here, this phenomenological approach is taken, with Section 3 setting out the list of properties. This is close to an 'easy' problem approach, and borrows some of its scaffolding processes, but the phenomenological starting point pulls the resulting explanations closer to consciousness.

Perhaps once enough of these properties of consciousness are explained scientifically, interest in the supposedly remaining base of consciousness, and the mind-body problem, will fade away (Seth forthcoming). In this perspective, the "real" problem of consciousness is to match up phenomenology with underlying neural mechanisms. Here, we will be agnostic about the fate of the hard problem – it might persist even after all interesting properties of consciousness are explained.

An alternative way to approach the problem of consciousness is to broaden the presence of consciousness to a much wider class of things than humans, or certain animals, or things with brains. This might be a metaphysical view, panpsychism, which sees this as the only escape from the hard problem (Nagel 1979). Or is might be a theory-driven, naturalistic view, such as the view that since consciousness is integrated information, and integrated information is ubiquitous, consciousness is ubiquitous (Tononi and Koch 2015; call this 'infopsychism'). Self-evidencing could go that route, too, and claim that every self-organising system that self-evidences is consciousness (call this 'biopsychism').

There is much interesting reasoning behind panpsychism, infopsychism and biopsychism, but they are unworkable. There is wide agreement that many of the things that are supposed to possess these basic and ubiquitous "proto"-consciousness properties are not conscious in anything like the sense we humans are conscious. But it is exceedingly hard to scale up from these kinds of basic consciousness-properties to our familiar kind of consciousness. Attempts to scale up seem to import some critical functional or structural elements (e.g., of brain processes) along the way, which in fact ends up doing all the explaining. Varieties of "pan"-consciousness are therefore most likely explanatorily otiose.

A further approach to the hard problem of consciousness is to argue that it is only an apparent problem. There are several routes here, including takes on predictive processing (Clark, Friston et al. 2019). These approaches are interesting but not decisive for the science of consciousness. Even if the metaphysical problem of consciousness were some kind of philosopher's chimera, we would still be confronted with the same formidable and perplexing (and intriguing) task of making scientific inroads on our understanding of the nature of consciousness.

Hence, the project undertaken here is neither pan- or biopsychist, and neither does it attempt to solve, or dissolve, the hard problem. It puts familiar properties of consciousness first and seeks to explain them in terms of a subclass of self-evidencing mechanisms.

3. Properties of consciousness to explain

The focus here will be on a set of properties of consciousness worth having, and worth explaining. This is partly driven by the considerations above. There is much interesting and important science to be done on these properties. It is not so important, for making conceptual and scientific progress, that there remain unexplained properties of consciousness, or that the metaphysical problem of consciousness remains.

This can be put with a different emphasis. Consider all the disorders and disturbances and alterations to consciousness that humans and some animals encounter. Disorders of consciousness such as the vegetative state, fugue states in epilepsy, contemplative states in wisdom traditions, psychedelic states, schizophrenia, delirium, drunkenness and dementia. All these and more interfere with various properties of conscious experience. They are important to understand for clinical, therapeutic and leisure reasons. But we can make scientific inroads on them by doing consciousness science: taking seriously that they are states of consciousness, and not mere cognitive or informational states. It seems irrelevant for this work whether the metaphysical problem is solved or not, or dissolved, or whether the smallest particles or systems or creatures have proto-conscious properties.

There is some room for pragmatic manoeuvring in selecting the properties of consciousness that it will be worth focusing on and explain within the self-evidencing framework. This is because what is regarded as explanatory is relative to context, including our prior information and interests and values. The approach adopted here is a fairly commonsense phenomenology, informed by background knowledge of conditions affecting consciousness, filtered through decades of philosophy of mind discussing these matters (see Hohwy 2013: Ch 10 for more on this approach). Finessing this set of phenomenological properties is then a discursive project. The following list is also fairly briefly indicated, and each aspect of consciousness has been treated extensively in substantial philosophical, psychological, and phenomenological bodies of literature.

Here is the list of properties of consciousness worth trying to explain:

Conscious perception. What everyone agrees is that *conscious perception* is a key aspect of consciousness. We wake up in the morning, and experience hits us in its full richness. Thoughts, feelings, moods, bodily sensations etc. are mixed in with perception of the world around us – sights, sounds, smells etc. – flow through our consciousness. Whatever is brought to bear on explaining consciousness should make central room for conscious perception. Importantly, perception includes not just the external world but also the inner world of bodily sensations.

Detached consciousness. Conscious experience is however not confined to perception of what is presented to our senses. Experience can be *detached* from the incoming stimuli. We dream during the night, our minds wander and we daydream while awake, we form imagery in different sensory modalities (visual, auditory, gustatory etc.), we believe things and think

about things that are not in front of us or don't even exist, like rewards we would like to obtain. All of this is part of our conscious phenomenology, so the capacity for detached consciousness is crucial for understanding it. The consciousness machinery should be able to run off-line.

Conscious unity and first-person perspective. Normally, perhaps always, the stream of consciousness has a certain *unity*. It does not split off in separate streams, and it is difficult to see what that would even be like (Bayne 2010). Consciousness seems to, as it were, congregate about you, flowing in one direction only. The unified conscious stream always or at least mostly has a *first-person perspective*, where experience is anchored in your perspective on the world, your body and your inner life. Perhaps this perspective is challenged in some conditions of depersonalisation and delusion (such as Cotard's delusion) but mostly consciousness is first-person (for discussion see, e.g., Metzinger 2004).

Conscious flow. Consciousness is usefully but metaphorically described as somehow a stream where conscious experience *flows* through our minds. One experience more or less seamlessly flows into the next. There is a rich and variable temporal structure to this flow, with the present occupied by a slightly temporally extended window of experience, straddling a bit of the past, the immediate now and into the future. This is sometimes called the 'specious present' because it is a construction of the mind that is in fact somewhat unmoored from the sensory input at this literal point in time. Music, for example, is experienced not one note at the time, but in a more temporally extended moving window of experience. Much of the vast literature here appeals to William James' early discussions (James 1890, Pockett 2003).

Next comes a set of phenomenological properties of consciousness that are more subtle, more hidden behind immediate perceptual appearances, and harder to articulate.

Sense-making and flow states. It makes sense to say that there is no intermediate possibility between being conscious versus not being conscious at all. There might be cases where it is difficult to ascertain if people or animals are conscious or not, or fully conscious or less than fully conscious, but we are not inclined to say that anyone's state is indeterminately conscious ((Bayne, Hohwy et al. 2016); there is discussion about this, as about any topic in consciousness science). But within a state of consciousness, there are ways or modes of being conscious, such as being attentive, having a mood colouring one's consciousness, being drugged or tired, etc. (Bayne and Hohwy 2016). One interesting distinction in these *background states* is between 'sense-making' and 'flow' states. *Flow states* are most well-known (Csikszentmihalyi 1990), often as something to aspire to, where one is fully immersed in what one is doing, with implications for both mood and time perception; flow states can be characterised by hyper-attention where not much else than the task-relevant experiences seem to enter consciousness. A useful contrast to flow is what may be called *sense-making states* where one is inclined to step back from the current experiential stream in order to make sense of it, trying out different interpretations before again establishing something closer to the flow state. (Some familiar sense-making states are described in cognitive science (Chater and Loewenstein 2016, Gershman 2019, Wojtowicz, Chater et al. 2021); note that it differs from the technical sense of 'sense-making' used in enactivist approaches, (e.g., Maturana and Varela 1987, Di Paolo 2005, De Jaegher and Di Paolo

2007)). Between entirely befuddled sense-making and fully immersed flow states lies a spectrum of states, with a patchwork of sense-making and flow-like states; likely most of our conscious waking hours are more in the middle of this spectrum.

Metacognition and introspection. Among the to-and-fro of everyday conscious experiences, there seems to be another element, where conscious experience is monitored. This is sometimes called *metacognition*, cognition about cognition. At times it is discussed in terms of higher-order thoughts, thoughts about thoughts, which may underlie our intuition that in experience we have awareness *of* mental states (Lau and Rosenthal 2011). It is also sometimes likened to forms of *introspection*. Introspection or metacognition can be an explicit task we undertake. For example, when the doctor or nurse asks where on a ten-point scale your pain is, or if someone, a police officer, asks how confident you are about something only you saw. It may be that there is always an element of metacognition hovering over our consciousness, even if we do not always engage in explicit introspection. It may be that this kind of internal monitoring infuses or gates what we are conscious of and how we are conscious of it.

Self. With the subjectivity and first-person perspective of consciousness comes the question what that subject or first-person is. What is *the self*? And how does the self factor into our conscious experience? This is again a rich and varied philosophical and psychological debate. Entering into more speculative phenomenological ground, it seems that the subjectivity and first-person perspective that perfuses conscious experience is not a mere abstract notion or geometrical point. The engagement of the self in consciousness comes with a perhaps vague conception of who that self is. What kind of person are you, what is your history and set of memories, what kind of body do you inhabit that determines your vantage point, what are your intentions and desires pointing into the future from here? Put differently, the specious present – the moving window of momentary experience – is itself suspended in and coloured by the past, present and future of the self.

Sense of being. Lastly, something perhaps a bit more abstruse. Consciousness seems to be accompanied by an inchoate *sense of being*. Jorge Luis Borges, in an early piece otherwise sceptical of the existence of a self, perhaps expresses it well, as "this consciousness of being, […] the immediate security of *here I am* that it breathes into us" (Borges 2000: 4). It is as if consciousness silently expresses the sense "I am here, now – I am." It is the feeling that having these experiences now matter deeply to me, not just in terms of what I need and think or feel and desire, but in terms of that fact that I am. It is hard to say how widespread the sense of being is across the phenomenology of different people. Perhaps it needs to be articulated before people become aware of it; perhaps it is so feeble that some would not consider it a core property of consciousness. It does tap into a rich tradition in the history of philosophy, which I cannot do justice to here. The reason for including it here is that it relates to a kind of pure subjectivity that sits close to the otherwise ineffable core of consciousness. In an attempt to seed further debate, it is then useful to put it into the context of self-evidencing.

That completes the list of phenomenological properties to be considered here. The properties seem central to the kind of consciousness that we normally enjoy and which, as put above, is worth caring about. It may be that none of them are necessary for

consciousness; perhaps alien conscious systems are conceivable that have none of these properties, but this would be a very different consciousness than ours – not consciousness much worth explaining. Different people may come up with different lists, or may have different formulations of more or less the same property.

4. Properties of self-evidencing matched with properties of consciousness

The list of phenomenological properties of consciousness constitutes the initial explanatory target for the predictive processing approach to consciousness. Even if, as argued, self-evidencing in itself is not sufficient for consciousness, there might be processes of self-evidencing possessed by some organisms and not others, which are explanatorily relevant for consciousness conceived in terms of these properties.

The notion of self-evidencing emerged in philosophy of science, within discussion of inference to the best explanation, or abduction (Hempel 1965, Lipton 2004). In an inference to the best explanation, some evidence is explained by some hypothesis – that best explains the evidence – which is then inferred as true or probable. Often the evidence *for* the hypothesis is just the very evidence it explains. This creates a benign circle where the hypothesis explains the evidence and the evidence is evidence for the hypothesis. The explanatory power of the hypothesis ensures the hypothesis has evidence for itself – it is self-evidencing.

In predictive processing, with the free energy principle, self-evidencing sits with notions of self-organisation and self-supervision (Hohwy 2020). There is no outside help or evidence to establish the hypothesis (conceived as the system's internal generative model) – the only evidence is the evidence the model can account for. The system needs to configure or organise its own internal states (conceived as the model's parameters and states) in order to make the evidence (or sensory input) as expected or unsurprising as possible. As it accomplishes this task it accumulates evidence for itself. The task is non-trivial because the environment is changing and can undergo unexpected state transitions calling for diverse and different surprise-minimising processes. It is then possible for different creatures to go about self-evidencing in different ways, and these different ways would be inscribed in their model. Some of these ways may then be more explanatorily relevant for consciousness than others. (This kind of reasoning about self-evidencing can also be applied to cognition, rather than consciousness, where the argument is that some but not all systems are cognitive, in virtue of some of their specific self-evidencing processes (Corcoran, Pezzulo et al. 2020); the current argument extends this kind of reasoning to the case of consciousness). To the extent these ways of self-evidencing carve out a distinction among systems in nature, we also get a distinction among systems that are conscious and those that are not; or at least, as intimated earlier, systems that are conscious in ways worth caring about and worth explaining.

Below, each of the properties of consciousness are matched up with properties of self-evidencing, in ways that highlight their explanatory potential. Some of these self-evidencing properties are explained in more detail in the voluminous literature on the free energy principle and in various philosophical and introductory writings (Clark 2016, Buckley, Kim et al. 2017, Wiese and Metzinger 2017, Hohwy 2020, Seth forthcoming).

Conscious perception. A self-evidencing system may engage in *perceptual inference*. It may adjust the parameters of its hierarchical internal generative model in the light of weighted prediction errors, and thereby come to represent the causes of its sensory input. This begins to capture the first property, about conscious perception. More specifically, perceptual inference of this kind maps on to key phenomena and paradigms in consciousness science, such as binocular rivalry. Rivalry can indeed plausibly be explained in terms of this kind of predictive processing (Hohwy, Roepstorff et al. 2008, Weilnhammer, Stuke et al. 2017).

More recently, rivalry, and other phenomena relevant for consciousness, such as Troxler fading, have been modelled using predictive processing's key notion of *active inference* (Parr, Corcoran et al. 2019). Active inference relates to action and decision making. The key intuition is that a system can act in order to minimise its uncertainty, or do targeted testing of its predictions, namely by selective sampling of the sensory input expected under the model in question. Active self-evidencing in other words involves a kind of self-fulfilling prophesying. In its simplest form, active inference selectively samples input under the predictions of its model. For example, if it predicts a source of sounds, like a car, to move leftward, it may focus attention leftward, and increase gain on sensory input from that region. But there are other ways of engaging in active inference, which not all animals may possess. A system may be able to consider not just its current uncertainty, but also its expected uncertainty, under various scenarios. That is, the system may be able to consider counterfactual scenarios.

In this kind of counterfactual active inference, the system considers what the observations would be, were the system to enact one or another policy for action. Now active inference becomes a race between possible policies, or an inference to the policy that best augurs uncertainty reduction. For this to be a meaningful process, the system's internal causal model of the environment must be brought to bear on the question at hand. This includes the system's model of itself, in order to figure out how a given action performed by that system would give rise to certain observations. What makes a policy best is its lack of ambiguity, that is, the fidelity of its mapping from actions to outcomes, as well as the divergence of the outcomes given the action and the expected (or desired) state. Precise policies that tend to get the agent close to its expected (non-surprising) states tend to be inferred. When a policy is inferred, it is specified in terms of the observations it would give rise to. It is as if the system believes it is already executing the action, even before it actually is. That is to say, these still un-actualised observations take on the role of prediction error. Since the system is always trying to rid itself of prediction error, it will then selectively sample the sensorium, including its own bodily states (proprioception etc.), according to these inferred expectations. The system then enacts the inferred policy.

Some organisms may also make what has been labelled 'sophisticated' counterfactual inferences, where they explicitly represent the very beliefs that they would acquire, were they to enact a given policy, and then infer their policies under that counterfactual scenario (Friston, Da Costa et al. 2020). Sophisticated inference allows the agent to bring counterfactual models to bear on representations of themselves, yielding temporal depth and detachment from their current state. This is more genuinely counterfactual than standard active inference. It evaluates policies under a contrary-to-current-fact model of beliefs, rather than under the current, actual beliefs.

These more substantial counterfactual and sophisticated notions of active inference relate to properties of conscious perception in various ways. To begin, they imply that action is not exclusively in service of expected states in the sense of desires or rewarding states. It can also be in service of epistemic value, or uncertainty reduction per se. This captures ways of exploring one's environment in order to discover what it is like, even at a cost. Systems then need to learn about how to maintain a useful balance between acting for epistemic value and acting for utility. One consideration of course is that sometimes acting for utility can be premature, if the system has little inkling of how rewards might be distributed in the environment, and how complex it would be to get at them. In that case, the system is better off going around exploring, and only subsequently trying to target rewards. Acting for epistemic value implies that the system has a representation of its own uncertainty. It needs to be able to say "my knowledge is likely to have deteriorated, so I am unlikely to infer precise policies for utility, so I should be exploring over there in order to bed down my model again". This relates to properties of consciousness because agents imbued with such ability to selectively sample the environment in order to reduce uncertainty (that is, in order to self-evidence) can be shown to display quite substantial analogies to the perceptual dynamics of paradigms of conscious phenomena such as binocular rivalry and Troxler fading (Parr, Corcoran et al. 2019). There is thus some reason to think that conscious perception can be matched up with, and perhaps explained in some respects, by particular ways of self-evidencing that only some types of organisms possess.

Detached consciousness. Now consider detachment, the next of the selected properties of consciousness. Self-evidencing is rife with forms of detachment. At its most basic, self-evidencing systems are described and defined in terms of their enduring Markov blankets (Hohwy 2016, Kirchhoff, Parr et al. 2018, Palacios, Razi et al. 2020). Markov blankets are statistical descriptions of causal nets, in terms of Markov properties of conditional independence. They describe the boundary of organisms, identifying the external causes of sensory input to the blanket's sensory states, and the external effects of the changes driven by the blanket's active states. In between the sensory and active states are the internal, blanketed states. Markov blankets create a natural locus of detachment because, even though the internal states are causally connected to the external states, the conditional independence of the blanket means the behaviour of the internal states can be explained just in terms of the states of the blanket, independently of the external states. In other words, it is conceivable that there may be some detached interesting behaviour of the system for the understanding of which no appeal to the external states is needed. This is a promising start for addressing detachment.

Every self-organising system, even non-conscious ones, will have Markov blankets. So the detachment that comes just from possessing a blanket is not going to be directly informative about consciousness. But different types of systems can organise their internal states according to different processes, some of which speak more directly to conscious detachment. Consider that in addition to just revising models in light of prediction error and inferring policies, systems are well-served to engage in *model selection*. Systems that have a relatively long time-horizon over which they consider their expected uncertainty should select models that help them keep uncertainty in check. One process would be to select more simple models over more complex ones, as simple models (as long as they are not

overly simple) will accumulate less prediction error as time passes, since they are less likely to be fitted to noise. An interesting proposal is that some systems, such as humans and some other animals, dream as a result of engaging in model simplification, where unnecessary model parameters are pruned (Hobson and Friston 2012). This connects a canonical example of detached consciousness, dreaming to a type of self-evidencing that may not be seen in all self-evidencing systems (for further discussion of predictive processing and dreaming, Windt 2018).

Detachment can be found in other self-evidencing processes too. In prospective, counterfactual active inference of the kind explained above, there is a clear element of detachment, which speaks more directly to occurrent conscious perception (as compared to dreaming). Under active inference, the current sensory input is set aside and virtual, counterfactual observations take their place. This happens as the inferred policies are weighed against each other, and as the winning policy is enacted. The system assumes non-actual sensory input is actual, that is, it detaches from the actual input in favour of internally generated observations. It may well be that conscious perception then only occurs as those enacted policies are unfolding and the prediction errors are actually quashed through selective sampling. But even so there is an element of detachment that sits well with conscious perception. The striking fact about binocular rivalry and Troxler fading is that a large part of the sensory input is not allowed entry into consciousness (i.e., the suppressed eye's image or the peripheral stimuli). This is a kind of detachment from sensory input, which is explained well by the appeal to selective sampling in active inference accounts of these types of perceptual phenomena. To the extent consciousness is characterised by detachment, self-evidencing processes in some types of systems (and not others) may thus indeed be a good fit.

Conscious unity and first-person perspective. Next comes the unity and first-person perspective in the flow of the conscious stream. Self-evidencing does not make much sense in a wholly inactive system that is just more or less porously receiving input from the world. Such a system would just have some sensors and some internal states; it may engage in predictive coding, and a Markov blanket may be able to be drawn around it for some short period of time. But there would be little reason to suppose it can have a unified stream of experiences and a first-person perspective. Its internal processing could easily be disjointed, fragmented and distributed. Since we do not know the ultimate basis of consciousness it is not inconceivable that such a system is conscious, but it would be a consciousness much different to ours, without unity and a first-person perspective. If the system is allowed to act, and engage in active inference, then there is a better case to make for unity and a first-person perspective. For basic active inference, selective sampling according to predictions of sensory input would already yield a kind of singular perspective and focused stream. From a causal model perspective, the system is now a more distinct, unitary cause among other worldly causes, as it moves through the world and contributes to causal interactions. The sequence of sensory samples it selects are from different vantage points on the same world, and would provide a more or less continuous, more or less coherent, stream of states. This is a starting-point for unity and first-person perspective. However, it seems too rudimentary to capture what we have in mind for these as properties of consciousness (for discussion, see Hohwy 2013: Ch 10).

Systems that have counterfactual active inference seem better placed to have a unified stream and first-person perspective as we know it from our own consciousness (for a self-evidencing account along these lines, see Friston 2018). Now, in advance of acting (or as the system is enacting its current inferred policies) it is already looking ahead counterfactually at its next action. It marshals the relevant policies, those it has learnt in past experience are precise and unambiguous for the context at hand, and counterfactually plays them out to gauge the expected observations under them. This counterfactual exercise must represent the agent themselves, as a singular cause interacting with the world, to elicit the observations that would occur to itself as a policy unfolds. This first-person perspective would be more pronounced in sophisticated active inference, where the agent's own counterfactual beliefs are explicitly represented in iterated inference of policies. This seems to build in a more substantial first-person perspective onto mere agency, which anchors the observations under the inferred policy and creates a stream as the policy's control states (the individual movements making up the policy) are strung together to minimise the prediction error arising from inferring the policy.

Unity of the conscious stream also can also be accommodated because it is difficult to see how the fulfilment of the expected stream of observations would split into two or more, given the existence of one Markov blanketed system. If there were two streams, it would seem more natural to consider this two consciousnesses, enabled by two independent Markov blankets, rather than two streams in the same one consciousness. That is, from a causal perspective, the existence of two streams would look more like cell division or birth than a splitting of consciousness. This would perhaps still allow for fluid or intermediate cases, such as potentially some conjoined twins who may in some instances infer joint policies, and on other instances infer distinct policies. Hence, these properties of consciousness can feasibly be found in those creatures capable of engaging in active inference on the basis of their expected uncertainty (or free energy).

Conscious flow. The idea that within the stream of consciousness experiences somehow seamlessly *flow* one after another can also be captured (this sense of 'flow' differs from Csikszentmihalyi's notion of flow states, to be discussed later). It is tempting to think that the flow of conscious experience is driven by the flow of sensory input. As the world changes in front of one's eyes, so will the flow of consciousness. There is of course truth to this but change in sensory input is neither necessary or sufficient for change in flow. In inattentional blindness, quite saliently changing yet centrally fixated visual input is suppressed from consciousness. Dreaming is case where the Markov blanketed internal states of the brain somehow organise themselves to produce a flow of consciousness in the absence of input. Internally generated flow also occurs in binocular rivalry, where the different input to the two eyes (e.g., a face and a house) is kept constant and yet there is an alternating flow of conscious experience between the two images. This is of course an instance of the kind of detachment discussed earlier, but the question here is what explains the flow? In the studies analysing rivalry through predictive coding and active inference mentioned earlier (Hohwy, Roepstorff et al. 2008, Parr, Corcoran et al. 2019), the key trigger for alternation is the relaxation of priors. As the system settles into one inference (perceiving a face), the prior behind it will begin to decrease. The prior decreases due to a *volatility* belief to the effect that the probability of *state transitions* is increasing. In other words, it is unlikely that the world stays unchanged for very long, so the old prior will soon

be obsolete. A decreasing prior means that uncertainty is rising. Under active inference, uncertainty is accompanied by inference of a policy that is expected to reduce that uncertainty. In rivalry, this active inference is endogenous attention, increasing sensory gain for the input from the stimulus shown to the competing eye, which then comes to populate perception until its prior in turn decreases. In this way, the flow of conscious experience is driven, not by the world, but by how self-evidencing engages with our beliefs about the state transitions of the world; conscious flow is therefore surprisingly detached from the current sensory input (Hohwy, Paton et al. 2016).

Conscious flow can be conceived as a moving window (of typically approx. 200ms width; Pockett 2003), rather than as an extensionless point. This 'specious present' suggests that there is computation going on at the shortest time scales of the conscious stream, as snippets of already past input and soon-to-arrive input are included in the construction of conscious experience. However, it perhaps seems implausible that this computation is the laborious generation of sophisticated counterfactual outcomes for selective sampling. What is experienced in the specious present does not immediately *seem* counterfactual, rather it mostly seems relatively phenomenologically robust, even if laden with action possibilities. One potential way to accommodate this tension between robust presence and counterfactuality is through a combination of short time-scale *amortized* active inference that feeds into ongoing longer time-scale and truly counterfactual inference. In amortized active inference, policy-observation mappings from past inferences are re-used on new occasions, obviating the need to figure out parameters anew (Gershman and Goodman 2014, Millidge 2019, Friston, Da Costa et al. 2020). These efficient, rapid and ready-made inferences can then serve as the basis for more languid sophisticated active inference, as described above. This is an efficient strategy but only at short time-scales where the risk of underlying state transitions is low. Sophisticated inference could then describe an intriguing, layered phenomenology, encompassing a robust specious present imbued with a wider action-relevant, first-person perspective. Again, it may be that some systems, such as humans, have particular repertoires of amortized inferences, and particular capacities for iterating them in more flexible sophisticated active inference in cortical hierarchies of some depth. Other systems, who occupy highly stable niches, may more safely relegate much more processing to amortized inference.

Sense-making and flow states. The self-evidencing account of what drives conscious flow highlights the role of higher-order beliefs, such as beliefs about volatility, which governs certainty about our current perceptual or active inference. Behind the first-order representation thus lie continued, internal modelling of the ongoing state transitions, giving rise to beliefs about overall uncertainty, expressing the quality of one's ongoing self-evidencing. This can be used to throw light on *background states* of consciousness, such as the distinction, listed earlier, between attentionally focused and immersive *flow states*, and more searching, broad *sense-making states*.

The idea here is that immersion is facilitated by highly precise prediction error and unambiguous policies, made possible in part by low probability of volatile state transitions. The world has to not change behind one's back for flow states to occur. Conversely, we shift into sense-making mode, with a broader, more exploratory attentional profile, when we begin to believe that there are hidden causal interactions leading to state transitions, which

imminently will change our current grasp of the world, and prevent flow. Our sense of uncertainty about our self-evidencing may become so severe that we give up on the current model of the given situation – it may simply be accumulating too much uncertainty to be worth revising or guide selective sampling. This will then trigger (Bayesian) model selection, where we seek to more fundamentally make new sense of the situation. The system steps back and appeals to contextual cues and to the complexity and simplicity of the available models, before settling on one, and then commencing perceptual and active inference under it. Model selection is also part of self-evidencing, as one of the many ways the Markov blanketed system organises itself in preparation of active engagement with the world that can help it maximise evidence for itself. The modelling of the higher-order statistics of the world one represents thus drives processes that can help us understand some background states of consciousness.

These background states may be distinctively expressed in systems with particular repertoires of amortized inferences and capacities for iterated active inference. The low probability of volatile state transitions characteristic of flow states may afford an extended repertoire of amortized inferences, curbing the need for more open-ended counterfactual and self-involving processing. As state transitions being to emerge, the range of safe amortization shrinks and counterfactual, self-involving processing increasingly colours conscious experience, at least for systems with some hierarchical depth. Systems relying more on amortization will not experience the same character of background states.

Metacognition and introspection. There is much recent theoretical and empirical work on metacognition – cognition about cognition – which is being tied to consciousness in intriguing ways (Fleming 2021). Self-evidencing provides a convenient backdrop for metacognition. As described just above, in a changing, volatile world, some of the system's internal modelling must be given over to second-order statistical modelling of state transitions and uncertainty. This level of modelling is blind to the actual causes of sensory input and only feeds on the context-dependent fluctuations of uncertainty in the system. It has a direct impact on the processing of the first-order statistical, internal states it seeks to model, through controlling the gain on the prediction errors it "looks down at" (this aspect of self-evidencing is sometimes described in terms of precision optimisation and captures attentional mechanisms). This necessary aspect of self-evidencing provides the constant hum of internal monitoring, which could underwrite the property of consciousness considered here, about metacognitive, introspection-like, states of being 'conscious-of' one's own mental representations (Hohwy 2015). It may be that some systems, such as humans and some other animals, have developed ways to engage in this kind of metacognitive second-order statistics that bring metacognition closer to consciousness, such as the ability to actively and voluntarily select policies that dial up or dial down neural gain, irrespective of what the current uncertainty landscape is actually like (i.e., forcing attentional focus). Or, only some systems may be able to form explicit beliefs about their internal monitoring, and report them to doctors and police officers interested in judgements of confidence rather than reports of external states of affairs. More speculatively still, it may be that some systems can perform sophisticated inference upon their own metacognition, explicitly representing what their introspective beliefs would be under iterated counterfactual scenarios.

Self. Next there is *the self* and the question how it factors into our conscious experience. There are now several studies that seek to explain the self from the predictive processing perspective (Limanowski and Blankenburg 2013, Apps and Tsakiris 2014, Hohwy and Michael 2017, Perrykkad and Hohwy 2020). They tend to share the idea that the self is a model of inferred causes of sensory input. The idea here is that the agent themselves, their body and internal states are part of the causes of the sensory input the agent receives. The way we act in the world is influenced by what kind of self we are, and those actions bend back on the agent, determining what sensory input they receive. This input is prediction error, which is modelled in a self-evidencing, self-modelling process. Essentially, by modelling ourselves we will surprise ourselves less in the long-term average. This means that you do not know your self somehow directly, rather the self is inferred as a constellation of hidden causes that drives patterns in our behaviour. This also means the self is real, or as real as other causes of sensory input (Hohwy and Michael 2017). This kind of literal *self*-evidencing appears to be an attractive account of the nature of the self. It might also help explain the way self-awareness infuses our daily conscious experience. For systems capable of active inference on the basis of their expected uncertainty, running counterfactuals that map out actions-to-observations would continuously recruit and finesse the self-model. This is because extracting the sensory observations that would happen when executing a policy depends on the trajectory of one's body through space and time, and the internal causes (such as character traits) motivating the body's trajectory is informative for figuring this out. If, as argued earlier, active inference is central for conscious perception, and its flow in experience, then there would be traces of the self in all conscious perception. The self-trace in consciousness will be amplified in sophisticated active inference (as explained above), where the agent's own beliefs, under a counterfactual policy, are explicitly inferred. For systems with considerable hierarchical depth in their model, these explicitly represented beliefs about themselves could capture the types of relatively stable character traits we typically associate with our self.

Sense of being. Last on the list there was the *sense of being*. This aspect of phenomenology would be more fundamental and austere than the perfusion of experience with interoceptive beliefs, self-relating counterfactuals, or a structured, explicit self-model. It is not a representation of the causes of sensory input, neither in the world or in the system itself, so it is not directly related to the types of interoceptive inference thought to leverage sense of agency and affect as markers of homeostatic processes (Seth, Suzuki et al. 2012, Barrett 2016, Hesp, Smith et al. 2021). It would rather be an awareness of existence, tied to the very having of conscious experience. Being conscious helps us fathom our existence. To see a way of capturing this, consider that self-evidencing expresses the belief that "I exist", where the precision about this belief would rely on the rate of self-evidencing. If self-evidencing goes poorly, then the evidence for existence suffers – the belief is imprecise and flattens out. It may be that some systems, such as humans, can explicitly represent this precision of self-evidencing beliefs. To speculate further, sophisticated active inference at the highest level of abstraction could be associated with representation of expected "I exist" beliefs, resulting from iteration of counterfactual processing for precise policies. Though this belief is devoid of actual content apart from the rate of uncertainty reduction, it could give rise to a more or less confident sense of being. Obviously, these speculations provide only some first steps in using self-evidencing to account for this most subtle, but also, it seems, important aspect of conscious experience.

5. Concluding remarks

This completes the matching-up of types of self-evidencing with properties of consciousness. The selected list of properties can of course be debated, and the list of self-evidencing properties is also incomplete and is being expanded and finessed as the framework evolves. But it is a start on a project that can explain a constellation of salient properties of consciousness, and tie them together in a description of systems such as humans and some other animals, which self-evidence in distinct ways, which are unlikely to occur in every self-organising and self-evidencing system. Part of the motivation for this particular list of consciousness properties is that it picks out properties of a type of consciousness that we care about and that is worth having. It contrasts to types of possible consciousness that are very far removed from the human experience (e.g., as postulated in panpsychist or biopsychist positions). It is plausible that a case can be made that other properties should be added to the list of a consciousness worth having, for example about mood and emotions, or contemplative states of compassion and connectedness. This is welcome, for it will be an interesting research project to view such properties from the perspective of self-evidencing.

From a metaphysical perspective, instantiation of these self-evidencing processes in a system does not entail that the system is conscious. It is after all logically conceivable that there is a creature that self-evidences in just these ways, which is nevertheless devoid of consciousness, even if its internal representations have some of these properties. In other words, this is not a solution to the hard metaphysical mind-body problem.

Even if this account does not address the hard problem, it can be naturalistically and phenomenologically informative. It may be that we can better grasp the neurobiological grounding of consciousness by conceiving these salient phenomenological properties in terms of self-evidencing mechanisms (Hohwy and Seth 2020), and it may be that we gain a better handle on how to define and articulate properties of consciousness if we see them as manifestations of particular kinds of self-evidencing. The resulting position can be described as follows: the key explanandum is phenomenology, and a set of self-evidencing properties transpire as good explanations of why those properties are the way they are, including how they might not occur in creatures we do not consider conscious; hence it is an attempt at inferring self-evidencing as the best explanation of phenomenology. More work is obviously needed to establish that self-evidencing furnishes the best explanation. But it has promising beginnings of several classic “best-makers”: self-evidencing properties are unified with each other under the free energy principle, active inference in particular can explain in detail perceptual, cognitive and behavioural phenomena (Parr, Pezzulo et al. In print), and active inference integrates well with existing science on neurophysiology and anatomy, yielding promising process theories (Friston, FitzGerald et al. 2017).

The core claim is that distinct subset of types of self-evidencing explain phenomenological properties. The explanatory relevance has centred on computational self-evidencing mechanisms for generating the properties of content, confidence, context-dependence, nested and iterated dynamics of inference, and intrinsic hierarchical processes that can be discerned in phenomenology. There are significant questions in the philosophy of science about such naturalistic, bottom-up or reductive explanation. In philosophy of neuroscience,

there is useful debate about mechanistic explanation, mainly focused on cognitive mechanisms (Bechtel 2007, Craver 2007) with developments in particular around computational approaches and consciousness (Piccinini 2020). In phenomenological and physiological traditions, there is debate about how to frontload phenomenology into the explanatory project and what it would mean for there to be explanatorily substantial isomorphism between physiology and phenomenology (for discussion, drawing lines back to Kurt Koffka and Edwin Boring, see Feest 2021). An ongoing project here is to more closely marry the mechanistic and computational debates with the phenomenological debates, and thereby lay the ground for scientific projects that forge the explanatorily link for measurable properties of phenomenology (Ramstead, Hesp et al. 2021); a concrete example here could be the mentioned active inference account on Troxler fading and binocular rivalry (Parr, Corcoran et al. 2019).

The resulting picture of consciousness appears quite attractive. It ties consciousness closely to our basic sense of existence, our striving to make sense of the world and ourselves and all the hidden causes creating fluctuating uncertainty in our sensory input. It relies directly on our looping flows of perception and action, and the way they play out in us as we engage attentively with the world; and it portrays consciousness in contrast to other kinds of creatures who do not enjoy our kind of consciousness.

The picture of consciousness is also somewhat surprising, or even disturbing. It helps us explain the many different ways conscious experience detaches from current sensory input, which is an important aspect of the explanatory project. But it seems to make a virtue out of detachment. The theme is that detachment is driven by internal volatility beliefs, such that the content and flow of experience is determined by how the system believes it can best deal with changing rates of self-evidencing. This is less connected to the inflowing sensory input from causes in the world than we would perhaps normally think. Even in everyday perception of the things around us, there is mediation through active inference's penchant for relying on prediction error generated in a process of decidedly detached counterfactual processes. The detachment is not however wholly and disturbingly Cartesian or Kantian, because when the expected observations given a policy are then selectively sampled, the system needs the world to play its part. If the world does not afford to the system the expected uncertainty reduction, then self-evidencing suffers, and this would be reflected in conscious experience (for discussion of these topics, see Clark 2017, Hohwy 2017). Overall, it seems we can speak meaningfully of conscious self-evidencing, with a picture of our consciousness that is created and unfolds according to the system's own, internal self-evidencing beliefs and principles.

Acknowledgements
This research is supported by the Australian Research Council (DP160102770) and by Martin & Loreto Hosking's Three Springs Foundation. Thank you to Andrew Corcocan and the anonymous reviewers for helpful comments.

References

Apps, M. A. J. and M. Tsakiris (2014). "The free-energy self: A predictive coding account of self-recognition." Neuroscience & Biobehavioral Reviews **41**(0): 85-97.
Barrett, L. F. (2016). "The theory of constructed emotion: An active inference account of interoception and categorization." Social Cognitive and Affective Neuroscience.
Bayne, T. (2010). The Unity of Consciousness. Oxford, Oxford University Press.
Bayne, T. and J. Hohwy (2016). Modes of consciousness. Finding Consciousness: The neuroscience, ethics and law of severe brain damage. W. Sinnott-Armstrong. New York, Oxford University Press**:** 57-82.
Bayne, T., J. Hohwy and A. M. Owen (2016). "Are There Levels of Consciousness?" Trends in Cognitive Sciences **20**(6): 405-413.
Bechtel, W. (2007). Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. NY, Routledge.
Borges, J. L. (2000). The nothingness of personality. Selected Non-Fictions. E. Weinberger, Penguin Books**:** 3-9.
Buckley, C. L., C. S. Kim, S. McGregor and A. K. Seth (2017). "The free energy principle for action and perception: A mathematical review." Journal of Mathematical Psychology **81**: 55-79.
Carhart-Harris, R. L., R. Leech, P. J. Hellyer, M. Shanahan, A. Feilding, E. Tagliazucchi, D. R. Chialvo and D. Nutt (2014). "The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs." Frontiers in Human Neuroscience **8**.
Chalmers, D. (1995). "Facing up to the problem of consciousness." J. Conscious. Stud. **2**: 200.
Chalmers, D. (1995). "The Puzzle of Conscious Experience." Scientific American(December): 62-68.
Chalmers, D. (1996). The Conscious Mind. Harvard, Oxford University Press.
Chang, A. Y., M. Biehl, Y. Yu and R. Kanai (2020). "Information Closure Theory of Consciousness." Frontiers in Psychology.
Chater, N. and G. Loewenstein (2016). "The under-appreciated drive for sense-making." Journal of Economic Behavior & Organization **126**: 137-154.
Clark, A. (2016). Surfing uncertainty: Prediction, action, and the embodied mind. New York, Oxford University Press.
Clark, A. (2017). How to Knit Your Own Markov Blanket. Philosophy and Predictive Processing. T. K. Metzinger and W. Wiese. Frankfurt am Main, MIND Group.
Clark , A. (2019). "Consciousness as Generative Entanglement." Journal of Philosophy **116**(12): 645-662.
Clark, A., K. Friston and S. Wilkinson (2019). "Bayesing Qualia: consciousness as inference, not raw datum." Journal of Consciousness Studies.
Corcoran, A. W., G. Pezzulo and J. Hohwy (2020). "From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition." Biology & Philosophy **35**(3): 32.
Craver, C. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience, Oxford University Press, USA.
Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience, Harper & Row.
De Jaegher, H. and E. Di Paolo (2007). "Participatory sense-making." Phenomenology and the Cognitive Sciences **6**(4): 485-507.
Di Paolo, E. A. (2005). "Autopoiesis, Adaptivity, Teleology, Agency." Phenomenology and the Cognitive Sciences **4**(4): 429-452.

Dołęga, K. and J. E. Dewhurst (2020). "Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework." Synthese.
Feest, U. (2021). "Gestalt psychology, frontloading phenomenology, and psychophysics." Synthese **198**(9): 2153-2173.
Fleming, S. M. (2021). Know thyself: The new science of self-awareness, Basic Books.
Friston, K. (2010). "The free-energy principle: a unified brain theory?" Nat Rev Neurosci **11**(2): 127-138.
Friston, K. (2018). "Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?)." Frontiers in Psychology **9**(579).
Friston, K. (2019) "A free energy for a particular physics." DOI: arXiv:1906.10184.
Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck and G. Pezzulo (2017). "Active Inference: A Process Theory." Neural Comput **29**(1): 1-49.
Friston, K., W. Wiese and J. Hobson (2020). "Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism." Entropy **22**: 516.
Friston, K. J., L. Da Costa, D. Hafner, C. Hesp and T. Parr (2020) "Sophisticated inference." arXiv, arXiv:2006.04120.
Gershman, S. and N. Goodman (2014). Amortized Inference in Probabilistic Reasoning. Proceedings of the Annual Meeting of the Cognitive Science Society.
Gershman, S. J. (2019). "How to never be wrong." Psychonomic Bulletin & Review **26**(1): 13-28.
Hempel, C. G. (1965). Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. New York, Free Press.
Hesp, C., R. Smith, T. Parr, M. Allen, K. J. Friston and M. J. D. Ramstead (2021). "Deeply Felt Affect: The Emergence of Valence in Deep Active Inference." Neural Computation **33**(2): 398-446.
Hobson, J. A. and K. J. Friston (2012). "Waking and dreaming consciousness: Neurobiological and functional considerations." Progress in Neurobiology **98**(1): 82-98.
Hohwy, J. (2013). The Predictive Mind. Oxford, Oxford University Press.
Hohwy, J. (2015). Prediction error minimization, mental and developmental disorder, and statistical theories of consciousness. Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness. R. Gennaro. Cambridge, Mass., MIT Press**:** 293-324.
Hohwy, J. (2016). "The Self-Evidencing Brain." Noûs **50**(2): 259-285.
Hohwy, J. (2017). How to Entrain Your Evil Demon. Philosophy and Predictive Processing. T. K. Metzinger and W. Wiese. Frankfurt am Main, MIND Group.
Hohwy, J. (2020). "New directions in predictive processing." Mind & Language **35**(2): 209-223.
Hohwy, J. (2020). "Self-supervision, normativity and the free energy principle." Synthese.
Hohwy, J. and J. Michael (2017). Why should any body have a self? The Body and the Self, Revisited. F. de Vignemont and A. Alsmith, MIT Press.
Hohwy, J. and J. Michael (2017). Why would any body have a self? The Subject's matter: Self-consciousness and the body. F. Vignemont and A. Alsmith. Cambridge, Mass., MIT Press**:** 363-392.
Hohwy, J., B. Paton and C. Palmer (2016). "Distrusting the present." Phenomenology and the Cognitive Sciences **15**(3): 315-335.
Hohwy, J., A. Roepstorff and K. Friston (2008). "Predictive coding explains binocular rivalry: a review." Cognition **108**(3): 687-701.

Hohwy, J., A. Roepstorff and K. Friston (2008). "Predictive coding explains binocular rivalry: An epistemological review." Cognition **108**(3): 687-701.
Hohwy, J. and A. Seth (2020). "Predictive processing as a systematic basis for identifying the neural correlates of consciousness." Philosophy and the Mind Sciences **1**(II).
James, W. (1890). The principles of psychology. New York:, Holt.
Kanai, R., A. Chang, Y. Yu, I. Magrans de Abril, M. Biehl and N. Guttenberg (2019). "Information generation as a functional basis of consciousness." Neuroscience of Consciousness **2019**(1).
Kirchhoff, M., T. Parr, E. Palacios, K. Friston and J. Kiverstein (2018). "The Markov blankets of life: autonomy, active inference and the free energy principle." Journal of The Royal Society Interface **15**(138).
Lau, H. and D. Rosenthal (2011). "Empirical support for higher-order theories of conscious awareness." Trends in Cognitive Sciences **15**(8): 365-373.
Levine, J. (2001). Purple Haze: The Puzzle of Consciousness. Oxford, Oxford University Press.
Limanowski, J. and F. Blankenburg (2013). "Minimal Self-Models and the Free Energy Principle." Frontiers in Human Neuroscience **7**.
Lipton, P. (2004). Inference to the Best Explanation. London, Routledge.
Marchi, F. and J. Hohwy (2020). "The Intermediate Scope of Consciousness in the Predictive Mind." Erkenntnis.
Maturana, H. R. and F. J. Varela (1987). The tree of knowledge. Massachusetts, Shambhala Publications.
Metzinger, T. (2004). Being no one: The self-model theory of subjectivity, MIT Press (MA).
Millidge, B. (2019) "Deep Active Inference as Variational Policy Gradients." arXiv DOI: arXiv:1907.03876.
Nagel, T. (1974). "What is it like to be a bat?" Philosophical Review **83**: 435-450.
Nagel, T. (1979). Panpsychism. Mortal Questions. Cambridge, Cambridge University Press**:** 181–195.
Palacios, E. R., A. Razi, T. Parr, M. Kirchhoff and K. Friston (2020). "On Markov blankets and hierarchical self-organisation." Journal of Theoretical Biology **486**: 110089.
Parr, T., A. W. Corcoran, K. J. Friston and J. Hohwy (2019). "Perceptual awareness and active inference." Neuroscience of Consciousness **2019**(1).
Parr, T., G. Pezzulo and K. Friston (In print). Active inference: The free energy principle in mind, brain, and behaviour. Cambr. Mass., MIT Press.
Perrykkad, K. and J. Hohwy (2020). "Modelling Me, Modelling You: the Autistic Self." Review Journal of Autism and Developmental Disorders **7**: 1-31.
Piccinini, G. (2020). Neurocognitive mechanisms: Explaining biological cognition. Oxford, Oxford University Press.
Pockett, S. (2003). "How long is “now”? Phenomenology and the specious present." Phenomenology and the Cognitive Sciences **2**(1): 55-68.
Ramstead, M., W. Wiese, M. Miller and K. J. Friston (2020). "Deep neurophenomenology: An active inference account of some features of conscious experience and of their disturbance in major depressive disorder.".
Ramstead, M. J., C. Hesp, L. Sandved-Smith, J. Mago, M. Lifshitz, G. Pagnoni, R. R. Smith, G. Dumas, A. Lutz, K. Friston and A. Constant (2021). "From generative models to generative passages: A computational approach to (neuro)phenomenology."

Rudrauf, D., D. Bennequin, I. Granic, G. Landini, K. Friston and K. Williford (2017). "A mathematical model of embodied consciousness." Journal of Theoretical Biology **428**: 106-131.
Sandved Smith, L., C. Hesp, A. Lutz, J. Mattout, K. Friston and M. Ramstead (2020). "Towards a formal neurophenomenology of metacognition: modelling meta-awareness, mental action, and attentional control with deep active inference." PsyArXiv.
Seth, A. K. (forthcoming). Being You: Consciousness and the Beast Machine, Faber/Penguin.
Seth, A. K., K. Suzuki and H. D. Critchley (2012). "An interoceptive predictive coding model of conscious presence." Frontiers in Psychology **2**.
Tononi, G. and C. Koch (2015). Consciousness: here, there and everywhere?
Weilnhammer, V., H. Stuke, G. Hesselmann, P. Sterzer and K. Schmack (2017). "A predictive coding account of bistable perception - a model-based fMRI study." PLOS Computational Biology **13**(5): e1005536.
Whyte, C. J. (2019). "Integrating the global neuronal workspace into the framework of predictive processing: Towards a working hypothesis." Consciousness and Cognition **73**: 102763.
Whyte, C. J. and R. Smith (2020). "The Predictive Global Neuronal Workspace: A Formal Active Inference Model of Visual Consciousness." bioRxiv: 2020.2002.2011.944611.
Wiese, W. (2020). "The science of consciousness does not need another theory, it needs a minimal unifying model." Neuroscience of Consciousness **2020**(1).
Wiese, W. and K. Friston (2020). "The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation."
Wiese, W. and T. K. Metzinger (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. Philosophy and Predictive Processing. T. K. Metzinger and W. Wiese. Frankfurt am Main, MIND Group.
Williford, K., D. Bennequin, K. Friston and D. Rudrauf (2018). "The Projective Consciousness Model and Phenomenal Selfhood." Frontiers in Psychology **9**(2571).
Windt, J. M. (2018). "Predictive brains, dreaming selves, sleeping bodies: how the analysis of dream movement can inform a theory of self- and world-simulation in dreams." Synthese **195**(6): 2577-2625.
Wojtowicz, Z., N. Chater and G. Loewenstein (2021) "The Motivational Processes of Sense-Making." Available at SSRN.