On the intersection of the Free Energy Principle, The Hidden Spring, the Self-Model Theory, and Illusionism

In this article, I will present some literal “armchair” Philosophy of Mind that I like to indulge in when I’m not working, or spending time with loved ones. It’s important I start with this sort of watering down of your expectations because I’m not an expert in any of these subjects. I have, however, read a reasonable amount of the literature on each subject and I often find myself asking if there is some new ground one can stumble upon by taking a holistic view of certain theories or ideas about consciousness. That is the simple inspiration for this article. I will try to guide you through an interesting thought process that leads to the intersection of some relevant and established theories of consciousness. Here the goal is

to show how by considering these theories as different parts of a holistic theory we might discover something thought-provoking. I'll leave the verdict of how thought-provoking it is to you.

The Free Energy Principle — Why We Have Consciousness

Karl Friston's Free Energy Principle [1] is our starting point. This is not by accident at all. I think of the free energy principle more like a framework rather than a theory *of* consciousness. That is it describes a framework (a set of processes and rules) under which a self-regulating agential and (most likely?) conscious system operates. The free energy principle does not tell us what consciousness is but it does tell us why consciousness exists by describing a set of principles that gives rise to the existence of consciousness. I consider these principles a framework for consciousness. This view of the free energy principle is not new or controversial. Consider the concept of the *Markov blanket* in the free energy principle. The existence of the Markov blanket is what forces autonomous systems to model an internal world and then use this internal world model to make predictions about the external

world. Thus under the “framework” of the free energy principle, you can think of consciousness as nature's solution to the Markov blanket. So even though the free energy principle does not directly solve the hard problem of consciousness, it tells us why we have consciousness. In a highly chaotic, unpredictable, and hostile external world our consciousness allows us to make sense of the world and model the world, therefore allowing us to survive. Consciousness is the mechanism by which organisms minimize the free energy (entropy) in their environment. But that is not entirely the full story, there is a particular aspect of consciousness to whom emphasis must be given.

The Hidden Spring — The Importance of feelings, and its location in the brain

In Marc Solms' book “*The Hidden Spring: A Journey to the Source of Consciousness*” and in a more concise article [2] (amongst other publications), Marc Solms builds on existing work and presents a compelling view of how consciousness is connected to the free energy principle. Essentially consciousness enables us to minimize free energy or entropy within our environment. By subjectively

experiencing the world, we are equipped with the ability to navigate our environment and make choices that increase our chances of survival. Solms goes a step further though by highlighting a key element of consciousness that facilitates the relationship between a self-organizing conscious organism and the minimization of its free energy, that key element is *feelings*.

Feelings here refer to the very visceral form of consciousness that we are all familiar with. We feel pain, pleasure, fear, hunger, happiness, and sadness, we feel the sharp bites in our stomach when we have a stomach upset and we feel our muscles ache when we strain them. Our body is constantly relaying both external and internal feelings. It is these feelings that Solms argues allow us to navigate our environment, to minimize free energy, to survive. When you feel the sensation of hunger, you know that you have to eat, and if you ignore that feeling and refuse to eat then your chances of survival will start to diminish gradually. It is feelings that provide feedback from both our internal bodily environment and our external environment.

Now we know why consciousness and the more specific phenomenon of feelings are important, they help us minimize free energy, and they help us survive. However, we are still left with the hard problem unsolved, we know why we need consciousness but how does consciousness work precisely? While most modern science places emphasis on a cortical theory of consciousness, in The Hidden Spring, Marc Solms argues that *feelings* (and its associated subcortical structures) are in fact *all you need* to understand consciousness and solve the hard problem.

Solms starts his argument by discussing *the cortical fallacy* in chapter 3 of the book. Here he shows us that even though the everyday observation of our consciousness consists of perceptual images of events going on around us we should be careful not to conclude that the answer to the hard problem of consciousness resides solely in the cortex and its cortical processes such as visual, auditory, and language processing. He presents various medical cases to support this claim for instance in hydranencephalic patients where the brain has developed (in utero) without a cortex.

According to Solms, the fact these patients are still able to show emotions and react to external stimuli goes against the conclusion that they are vegetative and lack consciousness. He also highlights another example of decorticated mammals where the entire cortex has been completely removed surgically. Yet these mammals show complex goal-oriented behavior consistent with the presence of consciousness and feelings. One might still argue that these highlighted cases don't necessarily indicate the presence of consciousness and that these behaviors can result from subconscious processes, with no one actually *home* to experience anything subjectively. However, Solms also highlights cases where parts of the cortex that are typically touted as crucial for consciousness are injured and have been surgically removed such as the forebrain but fortunately, these patients still possessed the ability to communicate. In these cases, the patients are able to able to communicate a sense of selfhood and subjectivity, going against the claim that no one is home. In general, the goal here is to show that it is plausible to have consciousness without cortical structures and that we might be better suited to solving the hard problem of

consciousness by looking at brain regions that are absolutely essential to consciousness such as the subcortical structures typically involved in emotions and feelings.

This brings us squarely to the primary subject of The Hidden Spring which is that if we want to solve the hard problem of consciousness we have to understand the science of feelings. In chapter 11 Solms tackles Chalmer's "Hard Problem" head-on:

'[…] Why doesn't all this information-processing go on "in the dark", free of any inner feel?' In my view, the question only arose because Chalmers, following Crick, sought the function of consciousness in the wrong place. The fundamental form of consciousness is not something cognitive, like vision; rather, it is something affective. In that sense, and that sense alone, Chalmers was right to imply that consciousness is not a cognitive *function: the primary function of consciousness is not perceiving or remembering or comprehending but* feeling.

Solms then goes on to categorically state:

That is why I have focused the scientific arguments in this book upon feeling. *In order to solve the hard problem of consciousness, science needs to discern the laws governing the mental function of 'feeling'.*

In The Hidden Spring, Marc Solms does a good job of drawing our attention to the precise location *where* science might look to understand consciousness. According to Solms, consciousness is generated in the *upper brainstem* and involves the *reticular activating system (RAS)* and *periaqueductal gray (PAG)*. We should however pay close attention to the PAG. Here is how Solms describes the relationship between the RAS, PAG, and the forebrain:

The PAG is the final assembly point of all the affect circuits of the brain. So, whereas the forebrain is aroused by the reticular activating system, the PAG is aroused (as it were) by the forebrain.

We might think of the reticular activating system and PAG, respectively, as the origin and destination of forebrain arousal.

> In other words, there is a feedback circuitry between the RAS, the forebrain, and the PAG. But the PAG is at the center of that circuit, it receives all affective signaling from the cortex, musculoskeletal, and visceral nerves. According to Solms:

Putting it as baldly as I can: all affective circuits converge on the PAG, which is the main output centre for feelings and emotional behaviours.

> What this means is that through some unknown mechanism, the PAG is responsible for all feelings. All brain processes go to the PAG where feelings are "assigned" to them. A cortical theory of consciousness will claim that the PAG must still relay these feelings to cortical structures to bring them into conscious awareness. Given that there are patients without a cortex or missing key parts of the cortex still able to display behaviors that indicate that they can

indeed feel pain, emotions, etc. This implies that it is possible to build a theory of consciousness entirely focused on the workings of subcortical structures like the PAG. I will go as far as making the bold claim that based on the evidence discussed in The Hidden Spring, the PAG is a self-contained structure that is capable of conscious subjective experience.

Self-Model Theory — The PAG models itself

If we claim that the PAG has subjective experience, how does it achieve subjectivity? To answer this question we must formalize the problem of subjectivity.

One of the most important aspects of subjective experience is *selfhood*. When we talk about phenomenal consciousness we usually mean that there is a first-person experience of the world. This first-person in our subjective experience is what philosophers refer to as the phenomenal self. According to Metzinger [3], the phenomenal self endows our subjective experience with centeredness and perspectivalness, giving us a first-person

experience. The states that we experience appear to us as our own states or as happening to us, there is a strong conviction of ownership of our phenomenal experience. Thus, selfhood is directly connected to phenomenal experience and thus subjective experience. To understand subjective experience we must understand how selfhood arises and how subjectivity relates to selfhood. Metzinger's Self-Model Theory (SMT) [3] formalizes a mechanism for selfhood and subjectivity.

Let's briefly discuss SMT, according to Metzinger [3]:

SMT is predominantly a representational theory of consciousness, because it analyzes conscious states as representational states and conscious contents as representational contents.

We can think of a "representational state" as a certain (mental) content that references something, an object, an outcome, etc. SMT introduces a theoretical entity known as the phenomenal self-model (PSM). From here on I will refer to the phenomenal self-model as

simply the self-model. The self-model is a coherent self-representation, a consistent internal model of itself. Here is how Metzinger puts it:

….the self-model is an episodically active representational entity whose content is determined by the system's very own properties. [….] This type of analysis treats the self-conscious human being as a special type of information-processing system: the subjectively experienced content of the phenomenal self is the representational content of a currently active, dynamic data structure in the system's central nervous system.

Essentially, your subjective experience is a representational content of yourself or at least the currently active model of a part of yourself. How then do we experience this representational content as a first-person subjective experience? Metzinger explains:

The conscious representational states generated by the system are transparent, i.e., they no longer represent the very fact that they are

models on the level of their content. Consequently — and this is a phenomenological metaphor only — the system simply looks right "through" its very own representational structures, as if it were in direct and immediate contact with their content.

> It is this "transparency" that enables us to "feel" as though we are directly experiencing the world in first-person. Whereas we are modeling ourselves experiencing the world but misrepresent whatever is happening to our self-model as happening directly to us. Another analogy I like to use is to imagine you were staring at yourself in the mirror, the reflection in the mirror is merely a "model" of yourself. Now imagine we could superimpose some spider climbing on your reflection's face, but as you were watching this spider climb your reflection's face you suddenly felt as though the spider was climbing your real face! Similarly, while the brain maintains a self-model of itself, subjectivity is the mechanism by which this self-model misrepresents whatever is happening to itself as a first-person experience. In [3], Metzinger discusses the

“phantom limb” amongst other experiments, as evidence for a self-model.

It is evident that under the framework of SMT, the self-model in order to be an effective “self-model” must have access to all parts of itself. More importantly, it must have access to the necessary parts of itself needed to minimize free energy while navigating the world. Metzinger describes this “bodily self” as a functional anchor of the phenomenal space. The body is what grounds the self-model in its phenomenal space, in its experience of the world. Take note of this requirement of the bodily anchor for a self-model as we shall return to it shortly.

if we claim that the PAG is the seat of phenomenal consciousness, then we have to explain how the PAG implements subjectivity. I propose Metzinger’s self-model theory as an explanation for how. My conjecture is straightforward:

The PAG on its own implements the simplest self-model of the organisms body, one that is grounded only in visceral "raw" feelings, representations such as pain, pleasure etc. When we add a cortex on top of that we enhance this self-model with access to visual and auditory representations.

Returning back to our discussion of the "bodily self", I will argue that the PAG is indeed an excellent candidate for the sort of brain area that receives signals from "every" part of the body in order to maintain a self-model. Another area typically associated with this self-model is the prefrontal cortex but we now know that we can observe conscious behavior without the prefrontal cortex but lesions to the PAG will lead to death.

While I can't actually provide any evidence that the PAG receives signals from every part of the body to model a self-model, there are studies that do show that there are multiple pathways to the PAG from a large number of cortical and subcortical structures as well musculoskeletal, and visceral nerves which is quite unlike any other

brain area. However, these studies are focused more on how these pathways are substrates for fear or pain signaling [5].

Illusionism — The illusion of the subjectivity

Metzinger's SMT on its own does not strive to be classified as a theory of consciousness under the banner of a physicalist, radical, or conservative theory but due to its reliance on a functional representational framework for the conscious state one might think to classify it as conservative realism. This would be a mistake. SMT does not strive to make any claims about the "realness" of subjective conscious states. It claims that these states are representational and that due to their "transparency", the self-model achieves a sort of naive realism about its access to these states. The SMT presents the self-model as a functional representational entity but it does not claim that this entity has any intrinsic, ineffable quality. The entity and its subjective experience are simply representational. However, due to the transparency of these representational states, we appear to have direct access to these states and this is what misrepresents them as ineffable, intrinsic, etc. In essence, the representations in

SMT are *quasi-phenomenal*. Here is how Frankish [4] defines a *quasi-phenomenal* conscious state:

A quasi-phenomenal property is a non-phenomenal, physical property (perhaps a complex, gerrymandered one) that introspection typically misrepresents as phenomenal. For example, quasi-phenomenal redness is the physical property that typically triggers introspective representations of phenomenal redness.

The primary difference between a conservative realist theory and an *illusionist* theory is that the former believes that mental conscious states are phenomenal — real, qualitative, ineffable — while the latter believes that they are quasi-phenomenal. Given that the phenomenal properties of the SMT's conscious states are merely quasi-phenomenal, does that not make the SMT an illusionist theory?

Let's establish what an illusionist theory is fully. Keith Frankish [4] introduces illusionism:

[illusionism] by contrast [to realism] denies that the properties to which introspection is sensitive are qualitative: it is an illusion to think there are phenomenal properties at all. […] [illusionists] hold that the introspectable properties of experience are merely quasi-phenomenal ones.

By claiming that the PAG has subjective experience generated by maintaining a self-model whose conscious states are representational and quasi-phenomenal, we are inevitably bound to collapse into an illusionist theory of consciousness. Why is conservative realism not sufficient to describe the sort of consciousness we have been discussing up to this point? To be clear, we are talking about the consciousness of “feelings”, located in the periaqueductal gray whose subjectivity is defined by the self-model theory. By adopting the self-model theory we are forced to adopt a representational framework to describe how the PAG achieves subjectivity. However, we cannot make strong claims about the intrinsic nature of the representational states accessed by our self-model, the way a conservative realist might. This is because we

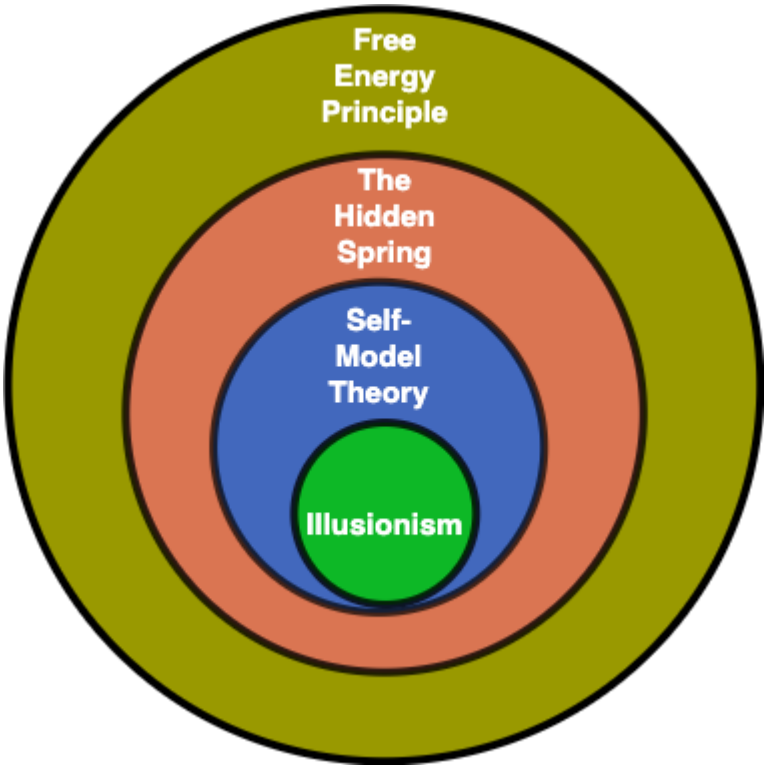
are also making the claim that these representations are transparent to the underlying neuronal dynamics that generate them. We are forced to admit that these representational states themselves are referential to some physical process and in this regard are quasi-phenomenal to those physical processes. Here is how Frankish describes this problem for conservative realism:

Indeed, one motive for advancing the strong illusionist position is to force conservative realists to face up to the challenge of articulating a concept of the phenomenal that is both stronger than that of quasi-phenomenality and weak enough to yield to conservative treatment.

If we assign phenomenal properties to the representations in SMT we must then describe how these representations are intrinsic and ineffable to a self-model keeping in mind that the self-model itself is a representation. This does not work. It is the same as asking how a virtual character feels virtual pain, the question is a tautology. Instead, we should ask why the virtual pain seems to be so “real” to

this virtual character. This is what Frankish [4] refers to as the *illusion problem of consciousness*. Thus, to understand how consciousness works in the system described in this article we should not attempt to solve the hard problem of consciousness but should instead solve the illusion problem of consciousness.

Conclusion



We have discussed how the concepts in the free energy principle, Marc Solms' The Hidden Spring, the self-model theory, and illusionism, complement each other to reveal a compelling

description of consciousness. Well, I promised to leave the decision of how compelling it is to you, so you be the judge of that. But no doubt, it is interesting if only as a mental experiment to see how these successful and well-acknowledged theories fit together. The overarching theme is that consciousness exists in its simplest form in a location in the brain known as the periaqueductal gray otherwise known as the PAG, this is the core thesis of The Hidden Spring. The PAG is responsible for feelings and it is these feelings that enable conscious organisms to minimize free energy. To understand how a brain area such as the PAG can be (or seem to be) phenomenally conscious we turn to the self-model theory (SMT) of subjectivity. SMT is chosen here because it accounts for how a structure like the PAG might model consciousness by aggregating signals from different parts of its embodied self. Finally, to explain why these aggregated signals appear so real and ineffable to the organism we turn to illusionism to show that one cannot prove the "realness" of a representational state to a representational system and that instead we should ask how and why that representational state seems to be so real to the representational system. I think the

answer to that question already lies between the lines of everything discussed here but it might suffice to dedicate another article to outrightly discuss this in detail.

[1] Friston et al; A Free Energy Principle for the brain: https://www.fil.ion.ucl.ac.uk/~karl/A%20free%20energy%20principle%20for%20the%20brain.pdf

[2] Solms; The Hard Problem of Consciousness and the Free Energy Principle:https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02714/full

[3] Metzinger; Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples: https://philpapers.org/archive/METEPF.pdf

[4] Frankish; Illusionism as a Theory of Consciousness: https://philpapers.org/rec/FRAIAA-4

[5] Vianna and Brandão: Anatomical connections of the periaqueductal gray: specific neural substrates for different kinds of fear: https://pubmed.ncbi.nlm.nih.gov/12715074/