Non-Experiential Evaluation

Jeremy Pober
Language, Mind, and Cognition Group
Centre of Philosophy
University of Lisbon

Forthcoming in *Philosophia*

0. Abstract

The framework Veit introduces for animal consciousness turns on finding and articulating its evolutionary origins. Veit argues that consciousness first evolved as affective experience in the Cambrian period. His argument centers around the plausible need of organisms in the Cambrian for a common currency of subjective valuation. I argue that such an adaptive pressure is unlikely to result in affective experience. I review other processes that instantiate common currencies of subjective valuation: unconscious (non-experiential) affect and the reward learning system and argue that each is likely to have evolved prior to affective experience.

1. Introduction

Veit's aim in *A Philosophy for the Science of Animal Consciousness* is to use evolutionary biology to create a standard by which to determine what animal species are—or are not—conscious. In a nutshell, the idea is to locate the evolutionary origin of consciousness in the 'tree of life' and determine whether an organism is a descendant of that initial species of conscious beings.

A central theme of Veit's is that consciousness should be defined in a way that "remove[s] humans from our centre of reference" (Veit 2023, 116). By this Veit means that the explanandum of his theory does not use aspects of human experience, especially those that are plausibly unique to human experience, as a benchmark for experience *simpliciter*. Thus when he talks about how consciousness evolved he is talking about something far less complex than human consciousness. Nonetheless, he clearly intends to talk about what Block (1995) calls *phenomenal* consciousness: in his words, he aims to answer "the question of what the subjective experiences of other animals are *like"* (Veit 2023, 24, italics in original).

Following Birch et al. (2020), Veit breaks down consciousness into different dimensions that putatively evolved separately. Veit locates the origins of consciousness with evaluative experience (alternatively: affective experience or hedonic valence), in the Cambrian period. Affective experience is the felt sense of pleasure, displeasure, pain, and other phenomena of the sort.

Per Veit's telling, newly evolved locomotive (and other motor) abilities endowed creatures with orders of magnitude more freedom than their ancestors possessed. Consequently, there was a need for efficient decision-making. And since any organism faces any number of factors to take into consideration when deciding how to act, the need for effective decision-making requires a *common currency* of evaluation, and evaluative experience is (or can act as) such a currency. It is here that I want to enter the discussion.

While it is true that affective experience can be a vehicle for a common currency of evaluation, it is not the only psychological system or process that can do so. Indeed, it already represents this common currency in tandem with other systems, notably the reward

(learning) system. And a closer examination of the relevant systems and processes will, I argue, demonstrate that other systems capable of representing a common currency of subjective valuation are extremely likely to have preceded affective experience evolutionarily.

Here, I discuss two possible systems. First, if we limit ourselves to affect, we find experimental evidence for *unconscious affect* (Berridge 1996; Berridge and Robinson 1998; Berridge and Winkielman 2003; Winkielman, Berridge and Wilbarger 2000). Unconscious affect bears the same relation to affective experience that blindsight bears to visual experience: that is, it has the same functional role as conscious affect, *sans* consciousness. Second, I argue that it is far likelier that the reward learning system evolved to meet the first need evolutionary need for a common currency. The reason is that affect can only represent a common currency occurrently: it cannot *store* information about past evaluations to use in the future. Nor can it change its evaluative responses to a type of stimulus in a principled way. Both of these functions—the former absolutely necessary for any survival benefit to come from evaluations, the latter plausibly responsible for the majority of their adaptiveness—are performed by the reward system.

I am not arguing Veit is necessarily wrong that affective experience evolved before perceptual consciousness. Rather, I am suggesting that he needs to make a different case for the evolution of affective experience. Crucially, the case needs to show the adaptive benefit of a *consciously felt* common currency, and not just a common currency *simpliciter*.

Two quick terminological notes before proceeding. I, like Veit, use 'affective experience' and 'evaluative experience' roughly interchangeably. However, I depart in using 'hedonic

valence' to refer to something potentially unconscious, i.e., non-experiential. Second, my use of 'representation' is deflationary: I am using it in the sense that Veit accepts when he says "affective decision-making may well be considered as representing two different states, 'good' and 'bad'" (Veit 2023, 66).

2. Background: Pathological Complexity in the Cambrian Era

Veit situates his case for the evolution of affective experience in an evolutionary framework that he introduces himself. The key concept for his framework is "pathological complexity" or the complexity of the organism's environment *and body* that matters to its adaptiveness. The concept offers a helpful way of thinking about adaptationism in that it recognizes the challenges an organism faces are i) cumulative, and ii) both external (from the environment) and internal (from the body). Both features play a role in Veit's account of the evolution of affective experience, to which I now turn.

In the Cambrian era, organisms first evolved multicellular bodies with distinctively animal traits like locomotive abilities. This led to an explosion of pathological complexity in both the environment and organisms' own bodies. Because animals could first move—if they were able to figure out how to coordinate their appendages and contort their centers—there was an exponential increase in what Veit terms the "organismal option-space" (Veit 2023, 79) or "degrees of freedom" (Ibid., 80).

As Veit notes (Ibid, 77), there were some limited locomotive organisms in the Pre-Cambrian period that lost their locomotive ability before becoming ancestors of today's

sponges. He reasonably takes the stance that the complex bodies with accompanying locomotive abilities became *mal*adaptive because they failed to evolve a way to deal with the very increase in pathological complexity that comes with having a body capable of locomotion.

What these Precambrian organisms needed but lacked, and what the Cambrian organisms evolved was "some form of informational bottlenecking … to deal with the problem of coordinating competing actions" (Ibid., 83). Something is needed to constrain the possible action space to those that might help the organism. What is needed is a way to quickly evaluate possible action options: "an evaluating system which enables the efficient deployment of the increase in [possible] behavioural complexity" (Ibid., 82). In other words, a common currency (Ibid., 75): a "value ranking on a common scale" (Ibid., 83).

What a common currency allows is the ranked ordering of all options in a set according to the organism's preferences. Even simple organisms with locomotive abilities need to consider factors like the location of food and predators as well as the cost of being in various environments (due to weather, temperature, etc.). Before we can consider two competing states of affairs (creatures with sensory consciousness like us conceive of our preferences in terms of states of affairs) that vary in all three respects, we have to figure out how to compare those factors to each other. Having a neurocognitively implemented common currency is what allows an organism to do so.

I find this picture—about evolutionary pressure for a common currency—plausible enough[1] to grant for the purposes of this discussion. Further, Veit is right that affective experience is a *candidate* to have evolved in response to this pressure at this time. It certainly meets what I take to be the central criteria for admissibility: first, it can instantiate a common currency of subjective valuation (Carruthers 2018). And second, because we know that affective experience is something present in the terrestrial tree of life, we know that developing it was available to evolution at some point in history.

But if these are (something like) the prerequisites for being a candidate, then affective experience isn't the only one. In what follows, I shall discuss two others: unconscious affect and the reward learning system. I argue that each are likely to have evolved before affective experience.

3. Unconscious Affect

On face, unconscious affect may seem like an oxymoron. That is, it may seem to be *part of the concept* of affect that it is felt. But we naturalists don't define psychological kinds by conceptual analysis alone: we do so by locating the referent of the kind term in the world (Griffiths 1997). If we individuate psychological kinds by functional role—whether or not we care about the physical mechanisms underlying or realizing those roles[2]—then if something

[1] Eventually, we will need more than plausibility—if Veit's account is to become the received one, this sort of evolutionary story will need empirical support—but that is not my concern here.

[2] Talk of functional roles individuating natural kinds, although the sort of thing found more in a previous generation (e.g., Putnam 1975; Fodor 1987), is compatible with and even arguably a restatement of the currently in vogue HPC theory of natural kinds (Boyd 1991), where the various aspects of a functional role—in particular its individual outputs—just are the properties in a a homeostatic property cluster.

has the same functional role as affective experience, then it *is* affect, even if it's not *experience*.

Berridge, Winkielman, and colleagues (Berridge 1996; Berridge and Robinson 1998; Berridge and Winkielman 2003; Winkielman, Berridge, and Wilbarger 2000) have demonstrated the existence of a state that has the exact functional role of affective experience *sans* conscious awareness. Berridge et al. call this state 'liking'—always in quotes or scare quotes—to both distinguish it from and note its similarity to consciously liking something. Winkielman, Berridge, and Wilbarger (2000) ran two experiments that support its existence. In the first experiment, subjects were subliminally presented with pictures of affectively expressive (e.g., positive as happy, negative as angry) facial expressions. Subjects were, after presentation of the subliminal stimulus along with its masks, asked to rate their subjective state and how it had changed since before the presentation: no significant change was reported. Subjects were then presented with a fruit drink and allowed to have as much as they liked. Subjects who reported being thirsty poured themselves more than controls after exposure to positively valenced faces, and less controls after exposure to negatively valenced faces. In other words, their behavior was influenced by the hedonic valence of the faces presented subliminally, but the subjects themselves reported no change in affective state. In the second experiment, subjects, instead of being presented with as much fruit drink as they wanted, were given a sip and asked how tasty they would rate it, and how much they paid for it. Again, these subjects reported no change in subjective experience but rated drinks higher (and worth more money) after exposure to positively valenced faces, and lower after exposure to negatively valenced ones.

Berridge and Winkielman ruled out the hypothesis that these behavioral changes were caused by affect-less yet valenced beliefs, noting that "subliminal facial expressions elicit genuine affective changes … including activation of the amygdala … and skin conductance responses" (Berridge and Winkielman 2003, 191). In other words, it doesn't just have the motivational aspect of affective experience's functional role.

The argument for unconscious affect having evolved prior to affective experience is straightforward. They have the same functional role, but one—the conscious one—is more complex and resource-demanding than the other. And *ceteris paribus*, the less resource-demanding adaptation will win out, all else being equal. Thus, even if some creatures in the Cambrian had mutations that allowed them experience, they wouldn't be our ancestors: they would be the creatures our ancestors with unconscious affect beat out.

Nonetheless, I don't intend to rest my case here, for two reasons. For I think *neither* version of affect would have been the first evolved trait capable of instantiating a common currency.

4. The Reward System

The reward system is essentially a reinforcement learning system that modifies an organism's behavior in ways that dispose it to obtain 'rewards' (Schroeder 2004). It is present in at least the common ancestor to all mammals. 'Reward' is understood in a technical sense here that is defined by neuroeconomists as a common currency of subjective valuation (Levy and Glimcher 2012)—the very type of common currency Veit and others (e.g., Carruthers 2018) take affect to carry.

As a reinforcement learning system, the reward system calculates the reward value of stimuli by instantiating an error prediction algorithm (Sutton and Barto 1998). In such an algorithm, the expected reward at a time t (predicted prior to t) is recorded and compared to (once t arrives) the actual reward obtained at t, which the system also records. When there is a mismatch a learning signal is generated (Schroeder 2004): when expected reward outweighs received reward, the learning signal lowers the computed reward value for that sort of stimulus in the future; when received outweighs expected reward, the opposite happens. The higher the reward value of a stimulus, the more, *ceteris paribus*, the organism is disposed to obtain it. The mechanisms behind this are detailed by Berridge and Robinson (Berridge 1996; 2012; Berridge and Robinson 1998; Robinson and Berridge 2008).

An important point for our purpose is that the reward system can operate unconsciously. We are clearly unconscious of the error prediction process aspect of reward learning: further, reward learning can occur *entirely* unconsciously.

First, there is evidence that our reward systems can respond to stimuli that do not reach our consciousness (Zedelius et al. 2014). Consider the following findings from Passiglione et al. (2009). Participants were asked to perform a task that they were told would be for monetary reward, but that the amount of reward would only be presented subliminally. Yet subjects showed greater brain activation in multiple parts of the ventral palladium associated with motivation, as well as greater autonomic arousal as measured by skin conductance response, in response to greater values. These results strongly suggest that reward value is perceived and processed unconsciously.

Second, as Schroeder (2004, 98) rightly points out, we are unaware of a great many of the *de dicto* contents of representations that the reward system responds to. In [REDACTED], I discuss the example of someone whose reward system is set up to respond to [the taste experience I had as a child while eating Milky Ways. The person, of course, thinks they simply want Milky Ways, but, like most adults, doesn't achieve the same taste experience with extremely sweet foods that they did as a child. Our reward systems respond to external temperatures of whose range we are only vaguely cognizant (and we are surely not aware of the metabolic factors that can change the temperature at which we are comfortable).

Crucially, all talk of the reward system 'updating,' 'raising,' or 'lowering' the values of stimuli—of reward (reinforcement) learning—implies that the reward system *stores* the reward value of various stimuli. This is something affective experience alone cannot do. For affect is by definition occurrent (this is true of unconscious affect as well as affective experience). It can't 'store' anything to be called up for later: 'storage' metaphors imply *standing*, rather than occurrent, states. Without a way to store reward values, the only way affective experience could maintain the same value for a kind of stimulus over time is if an organism felt the reward value of all kinds of stimuli they have encountered in their life histories. Organisms clearly do not do this (at least outside of the film *Everything Everywhere all at Once*). Thus, without stored values, there is no reason to think an evaluation of the same object in the same context at different times will lead to the same 'score'. But without such consistency, I am not sure that common currency-based evaluation really has any fitness value. To see why, it will be helpful to illustrate the phenomenology of Veit's proposed first conscious beings.

Veit's 'Benthamite' creature feels good or bad, much like Strawson's (1994) 'weather watchers' but unlike them has no idea of why they feel good or bad. At most, For the organism isn't conscious of what is impinging on it from the outside, rather, it is only conscious of a basic feeling of goodness or badness that is *caused by* those external stimuli. they will know a few proximate causes of their feelings, like being sated, and they will know which actions correlate with satiation, but they will not know why.

The creature's phenomenology is this way because, by hypothesis, it lacks sensory consciousness. As Veit (2023, 66) says, "[v]alence plausibly came into existence with a basic feeling of good and bad, without any *felt* sensory richness" (italics in original). Nonetheless, the creature needs something like *detection* abilities to get some sort of feedback from its environment.

For without such a process, there is little adaptive point in locomotion. Even with unconscious sensation, the adaptive value of a common currency is limited: without sensory consciousness, the common currency used by these organisms will not be a currency of objects (or states of affairs), but of behaviors that correlate with (unknown) objects. This is a much less efficient system from an adaptiveness perspective because behavioral means are always imperfectly correlated with ends—and that's in creatures who know what the ends are!

Without any external detection, all the organism can 'associate' with any desired result (be it affectively or unconsciously registered) is motor representations of its own movements, such as moving in a direction or grasping. But since predators and resources aren't correlated with things like 'being to the left of' or 'being above' an organism's body, those correlations

will be pretty limited to say the least. Locomotion only brings benefit when an organism has a clue of where to move.

Here is where storage comes into play. Having a clue of what movements to make involves having knowledge of what movements worked in the past. The benefit of feeling discomfort yesterday was that it led our little Cambrian friend to ingest some nutrients. If it can't 'remember' that, it might just flop around as a result of the same feeling of discomfort today. In short, there's little point in evaluating something as good or bad if you can't do anything about it.

But the evolutionary benefit of storage alone is rather mitigated without learning. Without learning, an organism is limited to the evaluations it was born with the dispositions to make. Such an organism would still have some benefit over one that could not make or store evaluations in a common currency, and it's plausible that such organisms evolved before those with reinforcement learning abilities. If this is the case, then the first (adaptive) solution to the need for a common currency wasn't the reward system itself but a progenitor of it that only stored reward values. Either way, though, a part of the total circuit representing a common currency *other than affective experience* evolved first.

5. Affect and Reward: Further Considerations

There is a potential objection to the argument I have just made which is worth addressing. One might think that for a reinforcement learning system to count as a reward system—or for its theoretical progenitor to count as storing information about rewards—it must be connected to affective experience. The basic idea is that what it is for a stimulus to

be rewarding is to be pleasurable, and that the notion of rewarding makes no sense outside of a capacity for conscious pleasure. If this is right, then affective experience necessarily evolved first, as any system extant prior to its introduction would necessarily be representing information about something other than reward.

There is some merit to this argument. In particular, it challenges the *stipulation* of a reward learning system by recognizing that reinforcement learning systems are differentiated by their inputs (and outputs). All reinforcement learning systems are formally equivalent insofar as they instantiate error prediction algorithms: thus, they must be distinguishable by their other properties, viz., their functional role. Here, inputs are the relevant dimension. Predictive processing accounts (e.g., Hohwy 2013) deploy reinforcement learning processes throughout the mind/brain: what separates a process predicting the identity (kind membership) of an object as opposed to its reward value is their functional roles. I agree with this intuition.[3]

However, the argument goes beyond that (correct) assumption, claiming that the relevant aspect of reward's functional role is some aspect of its relationship with affective experience, such that what is pleasant is necessarily rewarding.

One intuitive way of cashing this idea out, the 'direct' way, is to assert that a reward system needs affective states to trigger its learning signals for it to count as a reward system.

[3] Notably, Schroeder (2004) does *not* agree with this intuition: he takes the learning aspect of reward processing to be all that is necessary for it to be a reward system. For the reason I just mentioned, I disagree and am exploring this issue in a manuscript in preparation. I am also thankful to [REDACTED] for raising it independently in personal communication.

But Berridge and colleagues' work showed that reward processing can happen in the absence of affective *experience*. Moreover, they took it to indicate that the input to the reward system was from an unconscious evaluation mechanism of 'liking' rather than conscious affect.

But there is another, more subtle and 'indirect' way of making this argument, one which Veit is perhaps hinting at when he notes that "hedonic valence serv[es] as an impulse for efficient action selection *at the level of the organism*" (Veit 2023, 82, emphasis added). I understand Veit to be contrasting what are generally called 'personal' and 'subpersonal' levels (though perhaps 'organismal' and 'sub-organismal' would be more precise here), Roughly, personal level processes are those that are attributable to the whole organism and available to all of its systems (including, if applicable, conscious ones), and subpersonal processes are those that are constrained within their proprietary system and thus *not* available to consciousness (Burge 2010). Insofar as the inner workings of the reward system are not accessible to consciousness, it would be considered a subpersonal process.

But why should the impulse for action selection be at the personal level, accessible to all systems (including conscious ones)? Certainly, *decision-making* is by definition something done by a person. Yet that does not imply that all of the processes which influence decision-making are accessible to consciousness/at the level of the whole organism: unconscious influences and processes of which we are only conscious of the outputs are commonplace (see Kornblith [2012] for an excellent account of such processes and their role in reflection). The Winkielman, Berridge, and Wilbarger experiments showed that reward processing influences behavior even when we are unaware of it, suggesting that 'liking' is just such a process.

6. Conclusion

I have argued that the case Veit makes for affective experience having evolved in the Cambrian is wanting. In particular, it is a good case for some system that represents a common currency to have evolved, but this conclusion underdetermines a case for affective experience in particular to have evolved.

Not all hope is lost, however, and I will close with two optimistic reflections. First, I do think it is possible for Veit to make his case. What he needs to do is focus on the adaptive benefit of a common currency's being conscious, rather than the benefit of a common currency as such, otherwise, mutations resulting in simpler common currency instantiating systems will always win out. Speculatively, I believe something like a more functionalized version Veit's discussion of the role of affect in agency (on pp77-79) could be fruitful for theorizing about the role of specifically conscious evaluation.

Second, Veit emphasizes the "centrality of valence in our subjective experience" (Veit 2023, 65). In a sense, one of the aims of this book is to ground a sort of conceptual centrality of affect, over sensory *qualia*, in an evolutionary picture. Yet I wonder if this conceptual centrality need or even ought be so grounded. Perhaps affect is central to our subjective experience more generally in virtue of being *developmentally* but not evolutionarily prior.

Declarations

This work has been funded by grant no. 2023.09280.CEECIND from the Fundação para a Ciência e a Tecnologia, Portugal.

References

1. Berridge, K.C. (1996). "Food Reward: Brain Substrates of Wanting and Liking." *Neuroscience and Biobehavioral Reviews* 20:1-25.
2. Berridge, K.C. (2012). "From prediction error to incentive salience: mesolimbic computation of reward motivation." *European Journal of Neuroscience* 35:1124-1143.
3. Berridge, K.C. and Robinson, T.E. (1998). "What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?" *Brain Research Reviews* 28, 309-369.
4. Berridge, K.C. and Winkielman, P. (2003) "What is Unconscious Emotion? (The Case for Unconscious 'Liking')" *Cognition and Emotion* 17:181-211.
5. Birch, J., Schnell, A.K, and Clayton, N.S. (2020) "Dimensions of Animal Consciousness." *Trends in Cognitive Sciences* 24:789-801.
6. Boyd, R. (1991). "Realism, anti-foundationalism, and the enthusiasm for natural kinds." *Philosophical Studies* 61:127-148.
7. Burge, T. (2010). *Origins of Objectivity*. Oxford University Press.
8. Carruthers, P. (2018). "Valence and Value." *Philosophy and Phenomenological Research* 97:658-680.
9. Fodor, J.A. (1987). *Psychosemantics*. MIT Press.
10. Griffiths, P.E. (1997). *What Emotions Really Are*. University of Chicago Press.
11. Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.
12. Kornblith, H. (2012). *On Reflection*. Oxford University Press.
13. Levy, D.J., and Glimcher, P.W. (2012). "The root of all value: a neural common currency for choice." *Current opinion in neurobiology* 22:1027-1038.
14. Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). "How the brain translates money into force: a neuroimaging study of subliminal motivation." *Science*, 316:904-906.
15. Putnam, H. (1975). "The Meaning of 'Meaning'". *Minnesota Studies in the Philosophy of Science* 7:131-193.
16. Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.
17. Sutton, R. and Barto, A. (1998) *Reinforcement Learning: An Introduction*. MIT Press
18. Veit, W. (2023). *A Philosophy for the Science of Animal Consciousness*. Routledge.
19. Winkielman, P., Berridge, K.C., and Wilbarger, J.L. (2005). "Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value." *Personality and social psychology bulletin* 31:121-35.
20. Zedelius, C. M., Veling, H., Custers, R., Bijleveld, E., Chiew, K. S., and Aarts, H. (2014). "A new perspective on human reward research: How consciously and unconsciously

perceived reward information influences performance." *Cognitive, Affective, & Behavioral Neuroscience* 14:493-508.