Joseph Gottlieb[1]

Unlimited Associative Learning and the Theory-Light Approach to Non-human Consciousness

Abstract: *Birch (2022a) proposes a theory-light methodology for studying whether invertebrates have the capacity for (phenomenal) consciousness. The success of any such methodology turns on the positive markers it proposes, and whether they are genuinely ecumenical. After providing an account of what it is for a marker to be ecumenical, it is argued that one of the more influential set of markers offered — unlimited associative learning — clearly counts as positive evidence for consciousness on only a small handful of theories, rendering Birch's theory-light approach tacitly theory-heavy. What emerges is a broader lesson concerning what exactly is required for the theory-light framework to be successful.*

1. Methodologies for Studying Non-human Consciousness

Ascertaining when we should ascribe consciousness — that is, *phenomenal* consciousness — to non-humans is hard.[2] It is especially

Correspondence:
Email: joseph.gottlieb@gmail.com

---

1 Department of Philosophy, Texas Tech University, TX, USA. ORCID: 0000-0001-9014-8487.

2 Going forward, I largely drop 'phenomenal'. I assume for a subject S to have conscious awareness of x (experience of x) is for S to be in a mental state that represents x and that is itself conscious.

Journal of Consciousness Studies, **31**, No. 11–12, 2024, pp. 85–109
DOI: 10.53765/20512201.31.11.085

hard when the non-humans in question are at quite a distance from us phylogenetically, like invertebrates. Although many invertebrates, like bees, exhibit impressive behavioural capacities, it is hard not to wonder whether it all just runs in the dark, given just how different they are from us biologically and neurologically. This *distribution question* is amongst the most challenging in the study of consciousness.

In a recent influential and important paper, Birch (2022a) distinguishes three methodologies for studying non-human consciousness: *theory-neutral*, *theory-heavy*, and *theory-light*. On a theory-neutral approach (e.g. Tye, 2016), no assumptions about the nature of consciousness are made. Instead, we work with a list of 'behavioral, functional and anatomical similarities between humans and non-human animals and use arguments from analogy and inferences to the best explanation to settle disputes about consciousness' (Birch, 2022a, p. 134). The theory-heavy approach, by contrast, advises us to *start* with a theory of consciousness within the contemporary cognitive science of consciousness, such as global neuronal workspace theory (e.g. Dehaene, 2014), and then ask whether non-humans exhibit the relevant neurological or behavioural markers predicted by that theory. Birch argues that both the theory-neutral and theory-heavy approaches fail, and instead recommends a theory-light approach as a middle ground. On the theory-light approach, the only commitment is to a 'broad hypothesis' about the relationship between consciousness and various cognitive functions. Birch calls this the *facilitation hypothesis*: '[p]henomenally conscious perception of a stimulus facilitates, relative to unconscious perception, a cluster of cognitive abilities in relation to that stimulus' (2022a, p. 140). The theory-light approach is meant to be largely theory-invariant, with Birch declaring that 'all currently popular theories of consciousness are compatible with the facilitation hypothesis' (*ibid.*, p. 141).

My focus is on the viability of the theory-light approach. One point of potential criticism starts by asking about the facilitation hypothesis *itself* and whether *that* is truly theory-invariant. To ask this question is to ask whether it is by-and-large true, on most theories, that conscious perception facilitates cognitive abilities relative to unconscious perception. Maybe that's not so. For instance, a theory might tell us that consciousness is an evolutionary spandrel (e.g. Rosenthal, 2008; Robinson, Maley and Piccinini, 2015). This would not necessarily imply that consciousness is epiphenomenal, but it would imply that consciousness has little to no biological utility, which would make it surprising if it nonetheless facilitated various forms of cognition. Or perhaps unconscious perception doesn't exist, which indeed was arguably the default view throughout the history of philosophy. This would render the comparative claim at the heart of the facilitation hypothesis trivially false. Of course, proponents of such a view do not deny that there is non-conscious subpersonal sensory information processing — the visual system still carries out edge-detection, for instance — but we do not, for example, *see* stimuli non-consciously. Complex cognitive tasks performed by, say, blindsight subjects are not evidence of non-conscious visual perception, but evidence of degraded *conscious* visual perception that is left unreported due to conservative response criteria (Phillips, 2018; 2021).

Whatever merits these kinds of criticisms to the theory-light approach may have, the tack taken here is different. Common to these criticisms is that they look at the theory-light approach in abstraction from any of its implementations. A different kind of criticism of the theory-light approach — the kind pursued here — does not. Let an *implementation* of the theory-light approach be any specification of the cluster of cognitive capacities purportedly facilitated by conscious perception. A theory-light methodology will only be as good as the extent to which it provides us with cognitive capacities whose status as *proxies* or *markers* for consciousness is largely theory-invariant. Suppose we proposed 'schlearning' as a cognitive capacity facilitated by conscious perception. Even if it were true that all theories agreed that conscious perception of a stimulus facilitates, relative to non-conscious perception, *a* cluster of cognitive abilities in relation to that stimulus, that doesn't mean that there is any cross-theory agreement about what's in that cluster. If it turned out that only one theory of consciousness was compatible with schlearning being facilitated by conscious perception, then schlearning is in fact a theory-heavy marker — even if the facilitation hypothesis itself, silent as it is on what the relevant cognitive capacities are, isn't. If the facilitation hypothesis is to have any teeth as a middle-of-the-road methodology — a methodology that, while not wholly theory-free, is at least to some extent theory-invariant — we must uncover cognitive capacities whose facilitation by consciousness is something most could agree on.

And as we are considering non-human consciousness, the obvious choice (verbal report) is out.[3]

Happily, there may still yet be some hope. A prominent candidate implementation of the facilitation hypothesis focuses on various forms of *unlimited associative learning* (UAL), which is an advanced form of ordinary associative learning (Birch, Ginsburg and Jablonka, 2020). My thesis is in some ways polemical, but it comes with a methodological lesson.

The polemical component: focusing on trace conditioning as a form of UAL, I'll urge that trace conditioning clearly provides positive evidence for consciousness on only a small handful of theories. And what goes for trace conditioning, I claim, goes for other forms of UAL. A UAL-based rendition of the theory-light approach is thus tacitly theory-heavy.

This is the lesson that will emerge as we proceed: for a set of purported markers to be relevantly theory-light, it is not enough for them to simply fail to *mention* any particular theories. Being *intensionally* ecumenical is not enough. 'Schlearning is facilitated by consciousness' doesn't mention any theories. Neither does 'trace conditioning is facilitated by consciousness'. Yet that's no victory. What proponents of the theory-light approach should be after, I contend, is a marker (or set of markers) that is actually more or less neutral amongst theories — an implementation that is *extensionally* ecumenical.[4] This paper attempts the hard work of looking at what the various theories actually say, or may not say, about trace conditioning, along with the related tasks of uncovering the pertinent sense in which trace conditioning might be *evidence* of consciousness, and the conditions under which something counts towards a purported marker's ecumenicity.

3 As Bayne *et al.* (2024, p. 460) note, and *pace* Birch (2022a, p. 141), if we had no cross-theory agreement, and instead just had a bunch of different criteria (i.e. different capacities) whose facilitation by consciousness was each supported by a different theory, we would end up in *theory-heavy* territory, with the weighting of each criterion being determined by the support given to the particular theory that lies behind it. This may not require that we *fully* commit to any one theory, but the weighting process invariably requires us to rely on our *confidence* in those theories. See also Chalmers (2023).

4 Thanks to Colin Klein for help framing this point.

2. The Facilitation Hypothesis, Implemented

We begin with UAL and the theory-light methodology. This section has three tasks. The first: getting clear on what UAL is, and how it works (§2.1). This is essential for understanding *why* Birch thinks that UAL-associated abilities, like trace conditioning, provide strong evidence of consciousness: its connection to various psychological mechanisms, including attention, integration, and, most centrally, global accessibility. This will then segue to our second task, which is distilling down these various mechanisms to just one that encompasses the others, which will allow us to conduct our analysis more smoothly (§2.2). The third task will be to look more closely at the relevant notion of evidence in the theory-light framework (§2.3). It will be argued that the most apt construal of the evidential relation — what it is to say that E is evidence for H — will be one that is only partially probabilistic in nature.

2.1. Unlimited associative learning

Suppose that a subject S non-consciously perceives some stimulus *x* at *t*1 but then consciously perceives *x* at *t*2. The idea behind the facilitation hypothesis is that, holding all else fixed, at *t*2 S will be able to activate certain cognitive capacities with respect to *x* she couldn't prior. The claim under investigation is not just that conscious perception causes UAL, but that it is a perceptual state's *being conscious* that facilitates UAL.[5]

UAL is not simple associative learning. With UAL, conditioned stimuli can be compound, e.g. a red and green sour object. It can be *novel*, i.e. not reflex-eliciting or pre-associated with an unconditioned stimulus. Conditioned stimuli are subject to *second-order conditioning* (i.e. associated with some other, novel compound conditioned stimulus), to *trace conditioning* (i.e. can be temporally gapped with the unconditioned stimulus), and can have their associations with value *flexibly rewritten* (Birch, Ginsburg and Jablonka, 2020).

5 Birch (2022a, p. 145) says that facilitation is causal. However, that a conscious mental state M causes some cognitive behaviour B is ambiguous between (i) M's causing B in virtue of being conscious and (ii) M's causing B in virtue of some of its other properties. Of course, that M and B are correlated doesn't imply that M causes B, but even if M *does* cause B, M may not cause B *in virtue of* being conscious. We need (i) to feel confident that B is unlikely to occur outside of consciousness.

Building on foundational work by Ginsburg and Jablonka (2019), Birch, Ginsburg and Jablonka (2020) initially proposed UAL as an *evolutionary transition marker*: a set of properties that, when we find some creature with them, we thereby find evidence that an important turn in evolution — here, the turn from unconscious creatures to conscious ones — has come to fruition. Yet while UAL and the theory-light framework are mutually separable, there are connections. When discussing UAL's status as an evolutionary transition marker, Birch *et al.* tell us that:

> One of the main attractions of the UAL approach is that it tries to cut through the noise of disagreement to find a hidden consensus. The idea is that, for all the disagreement, there is a consensus that a list of hallmarks of consciousness are jointly sufficient (given the actual laws of nature) for consciousness. Much of the disagreement is about which, if any, of those jointly sufficient conditions are also necessary. (*ibid.*, pp. 18–19)

And later, UAL is said to be compatible with 'many specific theories about the nature of consciousness' (*ibid.*, p. 19). Since this is one of the main merits of a theory-light approach, it is not at all unreasonable to view UAL as a candidate ecumenical marker (*cf.* Halina, Harrison and Klein, 2022, p. 67). Indeed, when Birch formally introduced the theory-light framework, while UAL is not explicitly discussed, two forms of UAL — trace conditioning and rapid reversal learning — are offered as cognitive capacities facilitated by consciousness (2022a, p. 10).

There is, however, a difference that needs highlighting. Why might we think that conscious perception, relative to unconscious perception, facilitates UAL? The original contention was that UAL-associated abilities have as *necessary* conditions capacities that are themselves purportedly *jointly sufficient* for consciousness — global accessibility and broadcast, feature binding/unification, selective attention, intentionality, integration of information over time, an evaluative system, agency and embodiment, and registration of a self–other distinction (Figure 1). The implication is that UAL-associated abilities cannot be exercised outside of consciousness.

If UAL-associated abilities are genuinely *sufficient* for consciousness, this would be immensely useful in the search for non-human consciousness. Trace conditioning is directly measurable in a way that consciousness is not. Find trace conditioning, and you will be *guaranteed* to find consciousness.

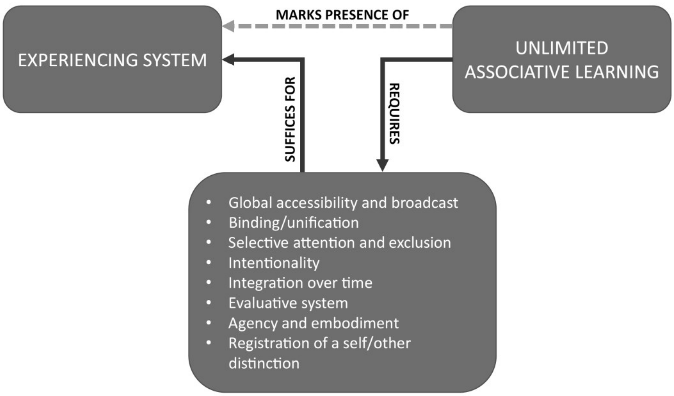

Figure 1. UAL as a transition marker for the evolutionary origin of consciousness (Figure 3 in Birch, Ginsburg and Jablonka, 2020, open access).

And this is where the difference lies: the theory-light framework does not traffic in sufficient conditions. Instead, the cognitive capacities in question provide *evidence* of consciousness, akin to the way symptoms relate to disease, with more symptoms providing more robust evidence (Birch, 2022a, p. 145). Exactly how we should understand 'evidence' here will be broached below (§2.3), but the key idea is that UAL-associated abilities are markers of consciousness in the sense that they are *reliable indicators* of consciousness (Allen and Trestman, 2020). This way of approaching the distribution question promises empirical tractability, but it does so without positing strict sufficient conditions. The upshot: when we ask whether UAL is really theory-light, we are asking whether UAL having an *evidential link* to consciousness is ecumenical or theory-invariant.

Before we press on, two points. First, and using Andrews' (2024) terminology, if UAL is a marker at all, it is not an *initial* marker, but a *derived* marker. Initial markers are 'pre-theoretical' in that 'they are not grounded in articulated theories about the nature of consciousness, philosophical solutions to the problem of other minds, or otherwise discursively learned through some intellectual practice' (*ibid.*, p. 7). Having a sensory or perceptual system is an initial marker. This is fine so far as it goes, but decades of experimental work have taught us that sensation and perception occur non-conciously.[6] Part of the value of

6 For a review of the empirical evidence, see Berger (2014).

an initial marker is how it gives way to more nuanced derived markers, which are inherently more theory-laden. What we are looking for is not perceptual systems as such, but perceptual systems whose representational outputs *meet certain conditions*. The facilitation hypothesis implicitly recognizes this point, since it turns on perceptual states not always being conscious. As does Birch, Ginsburg and Jablonka's (2020) understanding of UAL, which requires not just perceiving some stimuli, but that this perceptual state meets the conditions of being attended, globally accessible, and so on.

Second, it was noted earlier that while our focus is on trace conditioning, what goes for trace conditioning will go for other forms of UAL. We can make this simplification because UAL-associated abilities form a *natural cluster* that is ontogenetically correlated: they form a 'cluster of correlated abilities which are enabled by the overlapping underlying mechanisms' (*ibid.*, p. 13). Where we see one ability, we should expect to see others.[7] This suggests that if attention, global accessibility, integration, etc. really are required for, say, rapid reversal learning, they would also be required for trace conditioning. Thus, given these shared underlying mechanisms, in asking whether trace conditioning is an ecumenical marker, we are asking whether other forms of UAL are too.

2.2. Streamlining the trace conditioning–consciousness connection

In classical conditioning, a neutral conditioned stimulus (CS), like the ringing of a bell, is paired with a biologically significant unconditioned stimulus (US) like food. There is an unconditioned response (UR) — an unlearned reflex like salivation — to the US. When classical conditioning is successful, subjects will exhibit a conditioned response (CR) to the CS when the CS is presented alone. Trace conditioning has all these features, but with a twist: the CS and US do not temporally overlap. The CS begins and ends before the US is presented. For instance, you might hear a tone in your ear and, only after the tone ceases, feel a puff of air in your eye. The claim is that learning associations across this trace interval requires consciousness.

The central question for our purposes is *why* one might think that trace conditioning requires consciousness. The answer isn't: in humans, when there is no reportable awareness of the stimuli, trace

---

7 This 'cluster' talk also appears in Birch (2022a).

conditioning is interrupted.[8] We can't appeal to failure to report in non-humans when non-humans cannot report to begin with. And in any case, report-based paradigms are open to all sorts of confounds, and the evidence here is shaky, with trace conditioning being demonstrated in humans who are asleep (Arzi *et al.*, 2012).[9] What would be more useful is if we had some understanding of the psychological mechanisms that purportedly underly trace conditioning.

The good news is that, as we just saw, Birch, Ginsburg and Jablonka (2020) give it to us: trace conditioning, like any form of UAL, requires conditions that are themselves jointly sufficient for consciousness (see Figure 1). A tiresome approach here would be to go through each condition, examining what the individual theories have to say about them or their joint instantiation. It is not clear how profitable this would be, however, if only because most theories *don't* say anything in this regard, at least not for all of Birch *et al.*'s conditions. The conditions might seem, as Brown, LeDoux and Rosenthal (2021) point out, rather disparate.

That's only partially true, however. To see this, let's take a step back and begin with the claim that Birch and company ultimately want. Framing things in terms of *providing evidence* of consciousness rather than *being sufficient for* consciousness, we get:

(1) A subject S's performing trace conditioning with respect to stimulus *x* is evidence that S has conscious awareness of (has a conscious experience of; is conscious of) *x*.

The case for (1) begins with attention.[10] The use of distractors — and so the diversion of attention away from the target stimulus — is essential to disrupting trace conditioning. In the original Clark and Squire (1998) study, only those who noticed the temporal relation between the tone (CS) and air puff (US) anticipated the air puff by blinking in response to the tone. When the subject's attention was distracted by a movie, the processing of the tone–air puff contingency was inhibited (Droege *et al.*, 2021). As a result, the CR (the eyeblink)

8 See, for example, Clark and Squire (1998; 1999) and Clark, Mann and Squire (2002).

9 The study by Arzi *et al*. (2012) technically deployed *partial reinforcement* trace conditioning. That the reinforcement schedule was partial simply means that it was not employed on every trial.

10 By 'attention' I mean top-down (i.e. endogenous or goal-directed), object-based attention. All uses of 'attention' should be understood this way unless noted otherwise.

did not occur in response to the tone in anticipation of the delivery of the air puff.

Droege *et al.* (2021) remark on how this inhibitory effect was reproduced in a later study by Manns, Clark and Squire (2000) that asked participants to predict either the air puff or their own eyeblink after a tone. Trace conditioning did not occur in the eyeblink group; it was only successful in the air puff group. Again, this is presumably because predicting the eyeblink diverted attention away from the stimulus contingency; when the tone (CS) was attentionally inhibited, thus blocking an association with the later air puff (US), subjects failed to trace condition (Droege *et al.*, 2021). And this result seems to generalize to invertebrates, as the relationship between attentional inhibition and successful trace conditioning has been validated in insects (Paoli, Macri and Giurfa, 2023).

All of this suggests a trace conditioning–attention connection:

(2) If S performs trace conditioning with respect to *x*, then S is attending to *x*.

It would be a mistake to take (1) and (2) and argue that attention all on its own is strong evidence of consciousness. For that would imply that there is no non-conscious attention, and there is evidence that blindsight subjects can attend to the 'blind' portion of their visual field (their scotoma) yielding improved task performance (Kentridge, Heywood and Weiskrantz, 2004), and of non-conscious object-based attention in intact subjects (e.g. Norman, Heywood and Kentridge, 2013). While attention *is* one of the conditions listed by Birch *et al.* as necessary for UAL — so (2) itself isn't objectionable — the actual idea is that attention, global accessibility and broadcast, intentionality, self–other registration, etc. are *jointly sufficient* for consciousness. And if we were to, say, somehow shut off one's capacity for intentionality — one's capacity to represent — trace conditioning would obviously fail.

Fortunately, and *contra* Brown, LeDoux and Rosenthal (2021), attention is a *fulcrum* amongst many of the various disparate-seeming conditions on Birch, Ginsburg and Jablonka's (2020) list — information integration, intentionality, and (especially) global accessibility. For top-down attention is often considered *necessary* — and in many views, assuming requisite bottom-up signal strength, also sufficient — for global accessibility (Dehaene *et al.*, 2006; Wu, 2014). And global accessibility is itself sufficient for information integration since the function of a global workspace is to act as a 'blackboard' or 'router' in which the information from various cognitive modules is integrated (and then transmitted to various consuming systems). So, if we were to instead tender a global access-to-consciousness connection like:

(3) S's having a globally accessed representation of *x* is evidence that S has conscious awareness of *x*,

we would then have a claim that embeds not just Birch, Ginsburg and Jablonka's (2020) representation condition, but also attention, global accessibility, and, given that, integration.[11] Moreover, it is independently plausible that there is a trace-conditioning-to-global-access connection:

(4) If S can perform trace conditioning with respect *x*, then S has a globally accessed representation of *x*,

in so far as the learning involved in trace conditioning requires encoding in working memory (which is just what it is to be 'in' the workspace).

Our aim has been to condense the various mechanisms that underlie trace conditioning so as to make assessing the ecumenicity of (1) more manageable. This has been achieved with (3). In other words: the ecumenicity of (1) is just a matter of the ecumenicity of (3). Trace conditioning is a derived behavioural marker, and having a global workspace, with integrated and broadcast representational inputs, is a derived neurofunctional marker.

So, going forward, we'll trade asking about (1) for asking about (3). Whereas (1) is intensionally ecumenical, it's extensional ecumenicity becomes less clear once the shift to (3) is made. We therefore need to look at what the different theories say about (3). However, before we can fruitfully do that, we need to pin down one more thing: what is meant by 'evidence' in (1), and thus (3).

2.3. Evidence (of consciousness)

An account of the evidential relation will tell us necessary and sufficient conditions under which claims of the form:

E is evidence for H

11 It also plausibly includes evaluative systems since this is standardly understood as one of the specialized processers in a global workspace (witness any diagram of the global workspace).

are true, where 'E' picks out the evidence and 'H' picks out the hypothesis. Obviously, the nature of the evidential relation cannot be settled here. Fortunately, although the nature of the evidential relation is not made explicit in the theory-light methodology, there are hints internal to the framework that point us towards a suitable construal.

The first assumption we will make is that the evidential relation is probabilistic. As noted earlier, Birch (2022a, p. 145) says that the cluster of cognitive capacities (e.g. UAL) and consciousness stand in an evidential relation that is analogous to the one that holds between a cluster of symptoms and an underlying disease. Probabilistic notions of evidence are used frequently in medical diagnosis.

Now, on a simple version of the probabilistic approach to evidence, E is evidence for H if and only if E makes H more probable. That is:

$$P(H|E) > P(H).$$

Here, 'is evidence that' simply points to *probability raising*. On this view, it would not be terribly surprising if both (1) and (3) were accepted by, or at least compatible with, a wide range of theories. The problem though is two-fold: first, such an account of evidence seems false, and second, the theory-light framework itself seems to require something stronger.

As stated, this view of evidence is famously undemanding. What we want is for it to be the case that evidence gives us *good reason to believe* H. In other words, E must make believing H more reasonable than believing ¬H. That is, we at least want:

$$P(H|E) > P(\neg H|E)$$

Or:

$$P(H|E) > 0.50$$

It seems, however, that even this isn't sufficient for E to be evidence of H; high probability is not sufficient for evidence. Consider this example inspired by Achinstein (2001, p. 145). Let E be 'Carmichael has regularly taken his wife's birth control pills', and H be 'Carmichael has not become pregnant'. P(H|E) is in fact *extremely* high, but that Carmichael has taken his wife's birth control pills is not evidence that he has not become pregnant. The problem, according to Achinstein, is that the evidence needs to have an *explanatory connection* to the hypothesis.

This is Achinstein's (*ibid.*, pp. 151ff.) fix. Where k is some threshold, if E is to be evidence of H, it must also be the case that:

P(There is an explanatory connection between H and E|H & E) > *k*.

This handles the birth control case, precisely because there is *no* probability of an explanatory link between Carmichael's taking birth control and his not becoming pregnant.

This tallies nicely with the facilitation hypothesis, as facilitation is itself an explanatory relationship. The idea is not simply that trace conditioning is correlated with consciousness; the idea is that it is *in virtue of* mental state M's being conscious that its subject can perform trace conditioning (with respect to what M represents).[12] Consciousness is what does the facilitating. As such, it only makes sense that we'd want the theory-light framework to employ a notion of evidence that takes explanation seriously.[13]

I won't belabour the various points of controversy around evidence, probability, and explanation further.[14] And my hope is that we can work with an intuitive notion of explanation, perhaps one that is at

12 Recall note 5.

13 One might object that, even if we grant that the evidential relation requires an explanatory connection, it needn't be the case that consciousness *itself* is doing the explaining. Perhaps instead trace conditioning is causally explained by the *grounds* of consciousness (thus making trace conditioning associated with consciousness by metaphysical necessity), or perhaps it is causally explained by whatever neurological structures, in the actual world, *realize* consciousness (thus making trace conditioning associated with consciousness by nomological necessity). Either option, one might think, would yield an explanatory connection, albeit one that is indirect. While these suggestions are not unattractive, it's not clear that they are workable. Firstly, if consciousness is multiply realizable, it's not obvious how the latter option would offer even an indirect explanation. In addition, several major theories of consciousness are purposely formulated in abstraction from the neurological details. For example, while higher-order theories are sometimes associated with the claim that the prefrontal cortex is the neurological basis of consciousness (Lau and Rosenthal, 2011), HO theories are actually neutral on the question of neurological realization and are articulated at the psychological level of description (Berger and Mylopoulos, 2024). And as for 'grounds', I don't know of *any* scientific theory of consciousness that employs such a recherche notion. Thanks to an anonymous referee for raising this point.

14 Note that the question of explanatory relevance also shows up elsewhere; arguably, this is what was at issue in the *problem of irrelevant conjunction* (i.e. the *tacking problem*) for hypothetico-deductive accounts of evidence (Glymour, 1980). Again, the claim isn't that Achinstein's account of evidence is true. It's just that it is defensible on its own terms and fits naturally with the theory-light framework. So, I venture that if one preferred some different account of evidence, it would have to be close enough to Achinstein's such that the results to follow wouldn't change. For a discussion of why we should appeal to the (objective) *probability* of there being an explanatory connection between E and H, see Achinstein (2001, pp. 150–1). For a response to Achinstein concerning the positive-relevance account, see Roush (2003).

least partially causal. So going forward, I'll interpret the claim that trace conditioning (and UAL) is evidence of consciousness like this:

(i) P(S is conscious of x | S performs trace conditioning with respect to x) > k.
(ii) P(There is an explanatory connection between [S's being conscious of x] and [S performing trace conditioning with respect to x] | [S is conscious of x] & [S performs trace conditioning with respect to x]) > k.

And (3) will become:

(3*) (i) P(S is conscious of x | S has a globally accessed representation of x) > k.
(ii) P(There is an explanatory connection between [S's being conscious of x] and [S's having a globally accessed representation of x] | [S is conscious of x] & [S has a globally accessed representation of x]) > k.[15]

Call (3*) the *access is evidence of consciousness* claim, or 'AEC'. The exact value of k won't matter for our purposes — Achinstein puts it at ½ — and in what follows, the action will centre around clause (ii) of AEC, not clause (i). The question now is: what do the different theories of consciousness say about AEC?

3. The Theories

The immediate challenge with this question is that there are *many* theories of consciousness, and not all theories are attempting to do the same thing. Thus, when assessing the ecumenicity of a marker, we need know which theories *count* to that end.

Here is an example of a theory that doesn't. Panpsychism is a theory of consciousness. It says that consciousness is fundamental and ubiquitous. It is, as far as I can tell, completely consistent with AEC and a trace conditioning–consciousness link. But this doesn't score trace conditioning (or UAL more generally) any 'points' in so far as its theory-lightness goes, because the question of whether panpsychism is true was *never at issue* in the debate over the (macro-level) distribution of consciousness. Panpsychism doesn't make any

15 Following Achinstein (2001, p. 150), clause (ii) of (3*) allows that the explanatory connection can cut in either direction — H to E, or E to H.

methodological difference to how we would figure out whether, say, bees have the capacity for conscious experience. Panpsychism is a thesis about consciousness's *metaphysical* nature. That's not the kind of' 'theory' at issue.[16]

This is the criterion for ecumenicity I propose:

> For any theory of consciousness T, T's being *consistent* (*inconsistent*) with AEC counts *towards* (*against*) the ecumenicity of trace conditioning as a marker of consciousness iff T articulates the difference between conscious and non-conscious mental states — in particular, the difference between conscious and non-conscious perceptual states.

Though framed in terms of AEC, *the ecumenicity test* is a general ecumenicity test, applicable to other candidate markers. It is also a fair test, as the facilitation hypothesis presupposes that there is a difference between conscious and non-conscious perception, and that the former facilitates various cognitive capacities relative to the latter. On the test, a theory can be consistent with AEC but fail to count *towards* ecumenicity, and it can be inconsistent with AEC but fail to count *against* ecumenicity.[17]

16 Doesn't panpsychism entail that *everything* is conscious, and thus that bees are? No, and no — or at least, not panpsychism as such. The panpsychist's ubiquity claim is for consciousness at the fundamental (e.g. micro-) level. Whether panpsychism has implications for the distribution of consciousness at the macro-level depends on its combination principles. Thus, the claim isn't that metaphysical theories *cannot* have relevance to the distribution question (some have argued that *physicalism* does; see Birch, 2022b; *cf.* Carruthers, 2019). The claim is only that metaphysical theories don't have *direct* relevance in a way that other kinds of theories clearly do.

17 The ecumenicity test focuses on theories of conscious versus non-conscious *mental states* because the facilitation hypothesis is formulated in terms of conscious perception (perception being one kind of mental state). However, I assume that such theories extend to conscious versus unconscious *systems*, which is of course what we ultimately care about in the distribution question. Recall note 2: what it is to be a conscious system — both in the intransitive sense and transitive sense of being a subject that has experiences of things — *just is* to be a system that harbours mental states which are themselves conscious. A wrinkle: Gough (2024), who shares something like this assumption, argues that nonetheless we could have the following situation: while a system's being conscious requires that it be in M-type states with feature F, it may be that we only get M-with-F when M is instantiated in a system with feature F'. The implication is that we can't just look for F when trying to ascertain whether a system (e.g. a non-human one) is conscious. We need to look for F' too. I won't attempt to adjudicate whether this is correct, but simply note that, at least for the theories discussed here, there isn't any obvious answer as to what the systemic feature F' may be; the theories focus on F (e.g. being in a global workspace, being HO-targeted, etc.).

Let's begin with some easy cases. First, we have global neuronal workspace theory (GNWT). GNWT is obviously consistent with AEC. On GNWT, a perceptual mental state M is conscious when and only when, and because, attentional amplification allows M to be cognitively accessed and brought 'into' the global workspace (Dehaene, 2014). When that happens, S is conscious of what M represents. So, score one for AEC.

Prinz's (2012) AIR theory is a close cousin of GNWT. Here, a perceptual mental state M is conscious if and only if M is an attended intermediate-level representation. AIR and GNWT agree that the function of attention is selection for working memory. On AIR theory, attention only makes M *accessible* to working memory, and in this way is *disposed* to be encoded in working memory. By contrast, on GNWT a mental state's being in the workspace (and so conscious on GNWT) requires actual encoding in working memory (Wu, 2014).[18] There is still an obvious explanatory relationship between global access and consciousness on AIR, even if the former is not necessary for latter. For *if* a representation *is* globally accessed, it's at least in part because it already meets the conditions for being conscious on AIR.

Should we score another for AEC? Prinz (2012, p. 335) himself categorizes AIR theory as a species of GNWT. Instead of saying that we have two theories compatible with AEC, we could just as easily say that we have is in fact two types of *one* theory (i.e. GNWT writ large) compatible with AEC.[19] This point is rather fussy though. How, then, does AEC fair when it comes to other theories?

This is complicated. Consider *recurrent processing theory*, or RPT (Lamme, 2003; 2020). On RPT, non-consciously seeing a stimulus and consciously seeing a stimulus consist of distinct stages in visual processing. While an initial feedforward sweep of activity through the hierarchy of visual areas is sufficient for certain visual operations like feature extraction, it is not enough for visual *experience*; for that, you also need recurrent processing. This is where signals are sent back from higher areas in the visual hierarchy to lower areas.

---

18 See also Mashour *et al.* (2020, p. 785).

19 Tye's (2000) PANIC theory — where conscious states are poised, abstract non-conceptual representations — can, like AIR, also be read as a dispositional version of GNWT. Thus, the case for the theory-invariantness of AEC is not really boosted by including Tye's theory either.

As classically conceived, RPT is inconsistent with AEC. This issue is not simply that RPT allows for *phenomenal overflow* (Block, 2011). That thesis says that perpetual consciousness is *richer* than cognitive access, which only implies that we can be conscious of something without it being cognitively accessed, and so without it being in the global workspace. Rather, the issue is that, as a 'local' theory of consciousness, RPT proceeds under the assumption of a real, and not merely conceptual, distinction between *access consciousness* and *phenomenal consciousness* (Block, 1995; *cf.* Lamme, 2003). The former is arguably nothing more than encoding in working memory (Block, 2011), which, as we saw, is itself tantamount to the 'access' of being in the global workspace (Wu, 2014; Mashour *et al.*, 2020). Therefore, on RPT, we should expect global access and consciousness to doubly dissociate. That doesn't necessarily mean that Block's (1995) 'superblindsighter' will show up all that often. But it does mean that access and consciousness won't stand in an explanatory relationship, and in either direction. Indeed, proponents of RPT often agree with GNWT that it is *attention* that explains access, and that attention occurs outside consciousness.[20]

Now, some versions of RPT, like Malach's (2021, p. 2) 'local ignition' theory, might seem to posit a closer connection between recurrent processing and global access, claiming that 'complex contents… will be naturally linked to [a] multiplicity of cortical areas and large-scale network activation patterns'. But at best Malach is simply allowing that global access is explanatorily relevant to changes in the *phenomenal character* of more complex experiences; it is not explanatorily relevant to what makes a perceptual state conscious (and thus an experience) to begin with.

Next, we have higher-order (HO) theories of consciousness (e.g. Rosenthal, 2005; Brown, Lau and LeDoux, 2019). HO theories are a big tent. On classical variants, a mental state M of a subject S is conscious if and only if S is *aware* of herself as being in M via a suitable HO state M*. On the higher-order thought (HOT) variant, M* is a HOT. On the higher-order perception (HOP) variant, M* is a HOP.[21] On the higher-order map variant (HOM), M* is a HOM, which

20 For further discussion, see Montemayor and Haladjian (2019).

21 On some readings of HOP theory, the HO state is akin to attention (Lycan, 2008). However, what's attended to is not some external stimulus — as is the case in trace conditioning — but a first-order mental state. Prior proponents of HOP theory now concede that attention is not a means of *higher-order* representation (Sauret and Lycan, 2014).

is a cartographic representation that includes both iconic and discursive components (Gottlieb, 2022). In each case, there is no explanatory connection between global access and consciousness, as there is encoding in working memory without HO awareness (Gambarota *et al.*, 2022) and HO awareness itself doesn't facilitate encoding in working memory.

Here is a concrete case. Grover *et al.* (2022) present evidence that *drosophila* (fruit flies) can perform visual trace conditioning with temporal gaps of up to 20 seconds.[22] Assume, as we have been, that Birch, Ginsburg and Jablonka (2020) are right that this would require fruit flies to have a global workspace. Even assuming this, it is extremely unlikely that HO theories would say that fruit flies are conscious. For instance, on HOT theories, fruit flies would not just need concepts of their mental states, but also *self-concepts*: HOTs represent their bearers *as themselves*. That's likewise true on the higher-order map variant of HO theory, since the component of the cartographic representation that refers to the subject is discursively formatted (Gottlieb, 2022). This isn't to say that, on HO theories, consciousness begins and ends with humans; far from it (Berger and Mylopoulos, 2024). Indeed, the mirror self-recognition test — often considered evidence for the possession of self-concepts and self-awareness — has been passed by cleaner fish (Kohda *et al.*, 2023). But there is no evidence at all that insects (including fruit flies) can entertain *de se* representations. Representing their bodies (as being located in space) is not enough.

That said, other HO theories might seem less demanding in this way. For instance, Lau's (2022) perceptual reality monitoring theory (PRMT) claims that a mental state M is conscious if and only if it is targeted by a non-conceptual 'discriminator' which outputs a pointer (a kind of symbol that refers to a memory address). Since PRMT treats this discriminator as the means through which perception impinges on cognition, PRMT plausibly allows that there is an explanatory connection between global access and consciousness, although one that cuts in the opposite direction of GNWT.

Integrated information theory (IIT) is complicated in its own way. As Birch (2022a, p. 141) observes, IIT is compatible with non-conscious perception in so far as it is compatible with positing that there is perceptual processing in regions of sub-maximal Φ. Thus, one

22 For discussion, see Paoli, Macri and Giurfa (2023).

could potentially understand IIT as telling us the difference between conscious and non-conscious mental states. In that way, one could potentially understand IIT as a theory of consciousness in the sense relevant to the ecumenicity test. However, according to Koch (2019, p. 121), on IIT, 'a nearly silent cortex may be the substrate of a pure experience, without any ongoing information processing'. As traditionally understood, IIT tells us that consciousness does not have any strict psychological function. If that's so, it's at best unclear how a mental state's *being conscious* can *facilitate* anything, or in any way explain downstream cognitive functions, by way of global access and encoding in working memory. Koch (2019, pp. 120ff.) himself talks about how many of our behaviours are driven by the non-conscious portions of our mental lives. So compatibly with AEC is questionable.[23] Here, IIT's incompatibility with AEC *counts* against ecumenicity, since again we are assuming IIT is a theory in the requisite sense.

Granted, not all forms of IIT are the same. Let's distinguish between strong and weak readings of IIT, or what Michel and Lau (2020) call 'fundamental' and 'empirical' readings of IIT. On the former, consciousness is *identical* to (maxima of) integrated information. But on the latter, proponents of IIT are only advancing a weaker claim: integrated information is simply a marker of consciousness that can be used to detect various kinds of conscious states. And Michel and Lau further suggest that, on this weaker, empirical reading, IIT might end up compatible with GNWT (*ibid.*). Thus, it might be that, whenever a representation of the relevant stimuli (e.g. the tone–air puff contingency) is globally accessed, we have the right sort of informational complexity empirical IIT takes to be indicative of consciousness. Doesn't that suggest that at least empirical IIT is compatible with AEC?

No. For one thing, as we have repeatedly seen, even if global access raises the probability of information integration (in IIT's sense), global access isn't yet evidence.[24] We still need an explanatory

23 We already know that IIT will reject the idea that S's having a globally accessed representation of *x* explains S's being conscious of *x*.

24 Talk of 'integrated information' on IIT stems from a different place than on GNWT (Brogaard, Chomanski and Gatzia, 2021, p. 755). At least on the original, stronger version of IIT, information integration in the first instance stems from the purported phenomenological axiom that our experiences are *irreducible* in certain ways (the idea being that what it is like to see a red apple is irreducible to what it is like to see redness

connection. For another, empirical IIT is not a 'theory' of consciousness in the relevant sense. It does not tell us what consciousness *is* or offer a necessary and sufficient condition on a mental state's being conscious rather than non-conscious. In this way, even if a weak version of IIT is consistent with AEC, its consistency does not count towards ecumenicity. Such consistency would not support the original motivation to 'cut through the noise' and find 'hidden consensus' (Birch, Ginsburg and Jablonka, 2020, p. 4) amongst theories which are competing accounts of the same thing (like GNWT and RPT).[25]

How about attention schema theory (AST)? This view might seem to do better given that, as we saw, attention is explanatorily central to global access. Still, there are problems. One is simple: AST is not a theory of the difference between conscious and non-conscious perception. Graziano (2019, p. 96) is up front about this when he says that 'AST does not explain how a brain generates a subjective experience [but] how a machine makes claims about itself', i.e. the claim that it is conscious. Saying that *we claim* to be conscious by modelling attention is compatible with AEC but *not* in a way that counts towards ecumenicity.

This construal of AST is buttressed, albeit somewhat paradoxically, by the fact that AST is often combined with *illusionism*, offering a debunking solution to the meta-problem of consciousness (Chalmers, 2018).[26] Illusionism says that although we seem to be phenomenally conscious, we aren't (and neither is anything else). That makes AST trivially incompatible with AEC, at least if the latter is framed in terms of *veridical* evidence; that something provides veridical

---

and what it is like to see an apple). By contrast, when it comes to information integration on GNWT, the idea is that, for any given cognitive task, encoding in working memory ensures that the requisite information from specialized modules is combined together in the right way. One cannot successfully select the red, round thing unless colour and shape information from perceptual subsystems are combined (for a similar example, see Brogaard, Chomanski and Gatzia, 2021, p. 769).

25 Note too that, on weak IIT, the correct quantitative measure of integrated information is an empirical matter (Michel and Lau, 2020, p. 4), one that is not yet decided; it is not, as on fundamental IIT, determined before investigation by axiomatic facts internal to 'theory' (see note 24). It could turn out that we find some other measure of integrated information that better fits the data about consciousness yet fails to correlate with global access. Thus, compatibility between empirical IIT and AEC (and GNWT) is at best a moving target.

26 The meta-problem 'is the problem of explaining why we think consciousness poses a hard problem' (Chalmers, 2018, p. 6), and debunking solutions do so without invoking consciousness.

evidence of consciousness entails that consciousness exists. One might fashion a realist version of AST that offers more than a theory of our introspective judgments but, even then, compatibility with AEC remains murky. AST tells us that 'consciousness' (or the illusion of such) and attention, though correlated, are, as on RPT, doubly dissociable (Webb and Graziano, 2015, pp. 6–7). If so, there can't be an explanatory relation, and so evidential relation, between them.

Finally, there is midbrain theory (Merker, 2007). This view has gained some traction in the study of insect consciousness given that the midbrain is an evolutionarily old and subcortical structure (Barron and Klein, 2016). Now, it's not obvious how to formulate midbrain theory as a theory of consciousness in the relevant sense, as opposed to an account of the *neural correlate* of consciousness. But here's a stab: a mental state M is conscious just in case it is suitably realized in the midbrain, or some structure that is functionally equivalent to the midbrain (e.g. mushroom bodies in insects). Like GNWT and IIT, midbrain theory emphasizes information integration, but of a specific kind: 'an integrated neural simulation of the agent in space', combining both interoceptive and exteroceptive information (*ibid.*, p. 4903). It might be, as Birch (2022a, p. 141) offers, that on midbrain theory consciousness facilitates *some* cognitive capacities. Again though, the question is whether the midbrain theorist will allow any explanatory connection between global access and consciousness. They don't — to take just one direction, in humans the cortically-based workspace does not cause consciousness — and, to my estimation, it's hard to see how they could.

4. Conclusion

The importance of the theory-light methodological framework is not limited to its scientific value. It also has ethical significance. It is overwhelmingly plausible that possessing the capacity for (certain determinates of) consciousness is sufficient for being a *welfare subject*. It is sufficient, that is, for being the kind of thing that can be non-instrumentally better or worse off. And being a welfare subject is sufficient for having moral standing. So, if something has the capacity for consciousness, it has moral standing.[27]

---

[27] As far as I can tell, what's more contentious is whether (phenomenal) consciousness is *necessary* for being a welfare subject (Bradford, 2024). There are many complications

Though I have argued that UAL is hardly a theory-light marker of consciousness, if the avoidance of moral catastrophe were our sole aim — just consider how many insects are killed per year in farming practices alone — then this criticism might be much ado about nothing. After all, the evidential bar for precautionary standards is (rightly) quite low, and maybe possessing a marker (e.g. trace conditioning) that would be credible evidence for consciousness on two theories of consciousness is perhaps enough to meet that (low) bar. Yet when operating outside of precautionary contexts and trying to settle, as unequivocally as possible, the purely scientific question of the distribution of consciousness, the evidential standards are different. If the evidential value of a positive marker is derived in part from the fact that different theories converge on (or are at least neutral about) it's indicative relationship to consciousness, then trace conditioning — and, by extension, other forms of UAL — won't cut much ice in the search for which things are conscious.

Now, strictly speaking none of this is an indictment of the theory-light approach itself. Maybe there are implementations that focus on cognitive capacities other than sophisticated forms of learning, to which most theories would be amenable. The methodological lesson that has emerged is that any other implementation must be subject to the same battery of tests for ecumenicity: it must be extensionally ecumenical, not just intensionally so. With that said, I'd be remiss if I did not simply register some personal scepticism: it is not at all clear what these capacities might be, especially as the experimental evidence that many if not most behaviours can be driven by non-conscious mentality continues to grow. At the end of the day, there may not be any credible markers. Such an observation constitutes an indirect case for a theory-heavy approach. Spelling out just how that would work, however, is a task for another day.

Acknowledgments

Many thanks to Jacob Berger, Bob Fischer, Colin Klein, Ali Rezaei, and two anonymous referees at *JCS* for helpful comments and discussion.

here, however. For a discussion on the relationship between moral status (so nearby the notion of welfare subjectivity) and the distribution question, see Gough (2024).

References

Achinstein, P. (2001) *The Book of Evidence*, Oxford: Oxford University Press.

Allen, C. & Trestman, M. (2020) Animal consciousness, in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)*, [Online], https://plato.stanford.edu/archives/win2020/entries/consciousness-animal/.

Andrews, K. (2024) 'All animals are conscious': Shifting the null hypothesis in consciousness science, *Mind & Language*, **39** (3), pp. 1–19. doi: 10.1111/mila.12498

Arzi, A., Shedlesky, L., Ben-Shaul, M., Nasser, K., Oksenberg, A., Hairston, I.S. & Sobel, N. (2012) Humans can learn new information during sleep, *Nature Neuroscience*, **15**, pp. 1460–1465. doi: 10.1038/nn.3193

Barron, A.B. & Klein, C. (2016) What insects can tell us about the origins of consciousness, *Proceedings of the National Academy of Sciences USA*, **113**, pp. 4900–4908. doi: 10.1073/pnas.1520084113

Bayne, T., Seth, A.K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S.M., Malach, R., Mattingley, J.B., Menon, D.K., Owen, A.M., Peters, M.A.K., Razi, A. & Mudrik, L. (2024) Tests for consciousness in humans and beyond, *Trends in Cognitive Sciences*, **29**. doi: 10.1016/j.tics.2024.01.010

Berger, J. (2014) Mental states, conscious and nonconscious, *Philosophy Compass*, **9** (6), pp. 392–401. doi: 10.1111/phc3.12140

Berger, J. & Mylopoulos, M. (2024) HOTT and heavy: Higher-order thought theory and the theory-heavy approach to animal consciousness, *Synthese*, **203**, art. 98. doi: 10.1007/s11229-024-04529-8

Block, N. (1995) On a confusion about a function of consciousness, *Brain and Behavioral Sciences*, **18** (2), pp. 227–247.

Block, N. (2011) Perceptual consciousness overflows cognitive access, *Trends in Cognitive Sciences*, **15** (12), pp. 567–575.

Birch, J. (2022a) The search for invertebrate consciousness, *Nous*, **56** (1), pp. 133–153.

Birch, J. (2022b) Materialism and the moral status of animals, *Philosophical Quarterly*, **72** (4), pp. 795–815.

Birch, J., Ginsburg, S. & Jablonka, E. (2020) Unlimited associative learning and the origins of consciousness: A primer and some predictions, *Biology & Philosophy*, **35** (6), pp. 1–23.

Bradford, G. (2024) Consciousness and welfare subjectivity, *Nous*, **57** (4), pp. 905–921. doi: 10.1111/nous.12434

Brogaard, B., Chomanski, B. & Gatzia, D.E. (2021) Consciousness and information integration, *Synthese*, **198**, pp. 763–792.

Brown, R., Lau, H. & LeDoux, J.E. (2019) Understanding the higher-order approach to consciousness, *Trends in Cognitive Sciences*, **23** (9), pp. 754–768.

Brown, R., LeDoux, J.E. & Ronsethal, D. (2021) The extra ingredient, *Biology & Philosophy*, **36**, art. 16. doi: 10.1007/s10539-021-09797-z

Carruthers, P. (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*, Oxford: Oxford University Press.

Chalmers, D.J. (2018) The meta-problem of consciousness, *Journal of Consciousness Studies*, **25** (9–10), pp. 6–61.

Chalmers, D.J. (2023) Could a large language model be conscious?, *arXiv*. doi: 10.48550/arXiv:2303.07103

Clark, R.E. & Squire, L.R. (1998) Classical conditioning and brain systems: The role of awareness, *Science*, **280**, pp. 77–81. doi: 10.1126/science.280.5360.77

Clark, R.E. & Squire, L.R. (1999) Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies, *Psychological Science*, **10**, pp. 14–18.

Clark, R.E., Mann, J.R. & Squire, L.R. (2002) Classical conditioning, awareness, and brain systems, *Trends in Cognitive Sciences*, **6**, pp. 524–531.

Dehaene, S. (2014) *Consciousness and the Brain: Deciphering How the Brain Encodes Our Thoughts*, New York: Viking Press.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006) Conscious, preconscious, and subliminal processing: A testable taxonomy, *Trends in Cognitive Sciences*, **10**, pp. 204–211. doi: 10.1016/j.tics.2006.03.007

Droege, P., Weiss, D.J., Schwob, N. & Braithwaite, V. (2021) Trace conditioning as a test for animal consciousness: A new approach, *Animal Cognition*, **24**, pp. 1299–1304.

Gambarota, F., Tsuchiya, N., Pastore, M., Di Polito, N. & Sessa, P. (2022) Unconscious visual working memory: A critical review and Bayesian meta-analysis, *Neuroscience and Biobehavioral Reviews*, **136**, art. 104618. doi: 10.1016/j.neubiorev.2022.104618

Ginsburg, S. & Jablonka, E. (2019) *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*, Cambridge, MA: MIT Press.

Glymour, C. (1980) *Theory and Evidence*, Princeton, NJ: Princeton University Press.

Gottlieb, J. (2022) The higher-order map theory of consciousness, *Australasian Journal of Philosophy*, **100** (1), pp. 131–148.

Gough, J. (2024) Phenomenal consciousness and moral status: Taking the moral option, *Philosophical Psychology*, pp. 1–28. doi: 10.1080/09515089.2024.233935

Graziano, M.S.A. (2019) We are machines that claim to be conscious, *Journal of Consciousness Studies*, **26** (9–10), pp. 95–104.

Grover, D., Chen, J.-Y., Xie, J., Li, J., Changeux, J.-P. & Greenspan, R.J. (2022) Differential mechanisms underlie trace and delay conditioning in Drosophila, *Nature*, **603**, pp. 302–308.

Halina, M., Harrison, D. & Kein, C. (2022) Evolutionary transition markers and the origins of consciousness, *Journal of Consciousness Studies*, **29** (3–4), pp. 62–77.

Kentridge, R.W., Heywood, C.A. & Weiskrantz, L. (2004) Spatial attention speeds discrimination without awareness in blindsight, *Neuropsychologia*, **42** (6), pp. 831–835. doi: 10.1016/j.neuropsychologia.2003.11.001

Koch, C. (2019) *The Feeling of Life Itself: Why Consciousness is Widespread But Can't Be Computed*, Cambridge, MA: MIT Press.

Kohda, M., Bshary, R., Kubo, N., Awata, S., Sowersby, W., Kawasaka, K., Kobayashi, T. & Sogawa, S. (2023) Cleaner fish recognize self in a mirror via self-face recognition like humans, *Proceedings of the National Academy of Science*, **120** (7), e2208420120.

Lamme, V.A.F. (2003) Why visual attention and awareness are different, *Trends in Cognitive Sciences*, **7**, pp. 12–18. doi: 10.1016/S1364-6613(02)00013-X

Lamme, V.A.F. (2020) Visual functions generate conscious seeing, *Frontiers in Psychology*, **11**, art. 83. doi: 10.3389/fpsyg.2020.0083

Lau, H. & Rosenthal, D.M. (2011) Empirical support for higher-order theories of conscious awareness, *Trends in Cognitive Sciences*, **15** (8), pp. 365–373.

Lau, H. (2022) *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*, New York: Oxford University Press.
Lycan, W.G. (2008) Phenomenal intentionalities, *American Philosophical Quarterly*, **45**, pp. 233–252.
Malach, R. (2021) Local neuronal relational structures underlying the contents of human conscious experience, *Neuroscience of Consciousness*, **2** (2021), niab028. doi: 10.1093/nc/niab028
Manns, J.R., Clark, R.E. & Squire, L.R. (2000) Parallel acquisition of awareness and trace eyeblink classical conditioning, *Learning & Memory*, **7**, pp. 267–272.
Mashour, G.A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. (2020) Conscious processing and the global neuronal workspace hypothesis, *Neuron*, **105** (5), pp. 776–798.
Merker, B. (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine, *Behavioral & Brain Science*, **30**, pp. 63–81.
Michel, M. & Lau, H. (2020) On the dangers of conflating strong and weak versions of a theory of consciousness, *Philosophy and the Mind Sciences*, **1** (II), art. 8.
Montemayor, C. & Haladjian, H.H. (2019) Recurrent processing theory versus global neuronal workspace theory, *Philosophical Transactions of the Royal Society B*, **374**, 20180517. doi: 10.1098/rstb.2018.0517
Norman, L.J., Heywood, C.A. & Kentridge, R.W. (2013) Object-based attention without awareness, *Psychological Science*, **24** (6), pp. 836–843. doi: 10.1177/0956797612461449
Paoli, M., Macri, C. & Giurfa, M. (2023) A cognitive account of trace conditioning in insects, *Current Opinion in Insect Science*, **57**, 101034.
Phillips, I. (2018) Unconscious perception reconsidered, *Analytic Philosophy*, **59** (4), pp. 471–514.
Phillips, I. (2021) Blindsight is qualitatively degraded conscious vision, *Psychological Review*, **128** (3), pp. 558–584.
Prinz, J. (2012) *The Conscious Brain: How Attention Engenders Experience*, New York: Oxford University Press.
Robinson, Z., Maley, C. & Piccinini, G. (2015) Is consciousness a spandrel?, *Journal of the American Philosophical Association*, **1** (2), pp. 365–383.
Rosenthal, D.M. (2005) *Consciousness and Mind*, Oxford: Clarendon Press.
Rosenthal, D.M. (2008) Consciousness and its function, *Neuropsychologia*, **46** (3), pp. 829–840.
Roush, S. (2003) Positive relevance defended, *Philosophy of Science*, **71** (1), pp. 110–116.
Sauret, W. & Lycan, W.G. (2014) Attention and internal monitoring: A farewell to HOP, *Analysis*, **74** (3), pp. 363–370.
Tye, M. (2000) *Consciousness, Color, and Content*, Cambridge, MA: MIT Press.
Tye, M. (2016) *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?*, New York: Oxford University Press.
Webb, T.W. & Graziano, M.S.A. (2015) The attention schema theory: A mechanistic account of subjective awareness, *Frontiers in Psychology*, **6**, art. 500. doi: 10.3389/fpsyg.2015.00500
Wu, W. (2014) *Attention*, New York: Routledge.

Paper received January 2024; revised June 2024.