Evolutionary Transition Markers and the Origins of Consciousness

Marta Halina, David Harrison, and Colin Klein

1. Introduction

In recent years there has been a renaissance in research on the evolutionary origins of consciousness. A central goal of this research programme is to provide insight into the phylogenetic distribution of subjective experience: is there something *it is like* to be a chimpanzee? A raven? Or how about a bee? Answers to these questions facilitate inquiry into further questions regarding animal welfare and rights (see "What is the ethical significance of consciousness?", this issue).

Arguably the fullest treatment of the evolutionary origins of consciousness over the past few years is that advanced by the biologists and philosophers Simona Ginsburg and Eva Jablonka (Bronfman et al. 2016a, 2016b; Ginsburg & Jablonka 2019, 2021). They argue that such a study is aided by the identification of an evolutionary transition marker, which picks out the presence or absence of a property, trait, or capacity in evolutionary history. The appropriate marker for consciousness, they suggest, is *Unlimited Associative Learning* (UAL). UAL is a form of associative learning that allows a system to learn about itself and the world in an open-ended, exploratory manner (Birch et al. 2020: 8). That is, an organism with the capacity for *limited* associative learning can engage in classical and operant/instrumental conditioning but cannot make compound multimodal discriminations and has a very limited capacity for cumulative learning (Ginsburg & Jablonka 2021). In contrast, organisms equipped with *unlimited* associative learning can learn associations between novel, compound stimuli; associate stimuli that are separated in time or no longer present ('temporal thickness'); and engage in second-order conditioning. These capacities then allow for an "open-ended accumulation of long chains of associative links during an animal's lifetime" (Birch et al. 2020: 13).

Intriguingly, UAL, and the strategy of evolutionary transition markers in general, is intended to be analogous to the transition to unlimited heredity (see Smith & Szathmary 1995; Ginsburg & Jablonka 2015, 2019), whereby we can determine the transition from non-life to life. Indeed, in a similar manner to how "it is *the transition to unlimited heredity* that identifies sustainable living entities", so too is *the transition to UAL* supposed to identify definitively conscious organisms (Ginsburg & Jablonka 2019: 27).

Following this, it is clear that much turns on whether the cluster of capacities—broadly unified under UAL—are reliably seen as correlated with the presence of consciousness. If there were also good evolutionary reasons to think that such capacities were present early in animal evolution, then we might use the presence of UAL in different organisms to get a handle on the phylogenetic distribution of consciousness. Conversely, if we had a handle on the present distribution of UAL, that might give us a handle on the evolutionary history of the linked capacities. Treating UAL as a transition marker would thus permit two difficult questions to be fruitfully merged.

Recently, Birch, Ginsburg, and Jablonka (2020) have furthered exactly this line of thinking. They have rearticulated the UAL framework in a way that focuses on UAL as an explicitly *epistemic* marker of consciousness. The goal is to give conditions that are sufficient for us to reasonably believe than an organism is conscious. As Birch et al. write, they aim to "find hidden consensus behind the apparent disagreement by identifying a list of capacities that consciousness researchers would generally regard as *jointly sufficient* for a system being an experiencing subject" (2020: 4). They contend that UAL is a suitable evolutionary transition marker because it precisely requires this consensus list of capacities (see Section 2).

This epistemic project can be contrasted with a *metaphysical* project in which one identifies mechanisms, cognitive architectures, and enabling systems that support or give rise to consciousness. Elsewhere, Jonathan Birch (2020) rejects such an approach as 'theory-heavy', and the framework presented in Birch, Ginsburg, and Jablonka (2020) appears to eschew discussion or commitment to the specific mechanisms that might ground this project. Of course, such an eschewal comes with much to recommend it: for one, merely being able to delineate the set of definitively conscious organisms would itself be a substantial scientific advance. For another, figuring out what makes something conscious is not easy. So, starting with conscious organisms and working backwards has a lot to recommend it.

However, and as we will argue below, we contend that there is a risk to this strategy as well. Facts that suffice to show that something is the case need not *explain* why that thing is the case, not even a little. This is a well-known lesson from the philosophy of scientific explanation: the barometer may be an excellent marker for the approaching storm, but it does not explain the storm's occurrence (Salmon 1998). A good marker brings reliable connection, but not necessarily explanatory purchase.

Indeed, one might worry that the demands on a set of epistemically sufficient conditions are somewhat in tension with the demands of metaphysically sufficient ones. This is true even if we're trying to find the most defensible minimal sets. That is what we will argue. We consider two related problems here. First, metaphysically necessary and sufficient conditions need to be sensitive to edge cases: they need to include all of the, and only, conscious organisms, no matter how vague or strange. Good epistemic conditions, by contrast, favour all-things-considered accuracy. Hence, epistemically sufficient conditions might rightly shun edge cases in favour of increased clarity. But this also means that they are often quiet precisely when we need them to do work. Second, epistemic conditions do not lend themselves naturally to minimal sets: to use Matteo Mameli's (2008) distinction, they naturally pick out *clutters* rather than *clusters*. Hence, while they may usefully pick out a set about which to theorise, they don't really tell us much about why that set was picked out. We conclude by comparing the predominantly epistemic approach taken by Birch, Ginsbrug, and Jablonka (2020) with previous articulations of UAL as a transition marker (e.g., Bronfman et al. 2016). The latter is best characterised as 'theory-heavy' and we think it is more promising as an approach to investigating the origins of consciousness.

2. The Unlimited Associate Learning Framework

The core idea behind Birch, Ginsburg, and Jablonka's framework is that UAL serves as a 'transition marker' for the origin of consciousness. As they write, "A transition marker is a property such that, when we find evidence of it, we have evidence that the major evolutionary transition in which we are interested has gone to completion" (2020: 2). In this case, UAL indicates just such a move to completion in conscious animals. The reason for this is that UAL requires capacities that are themselves sufficient for consciousness. These capacities include global accessibility and broadcast, feature binding/unification, selective attention and exclusion, intentionality, integration of information over time, an evaluative system, agency, and embodiment, and registration of a self-other distinction (see Figure 1).

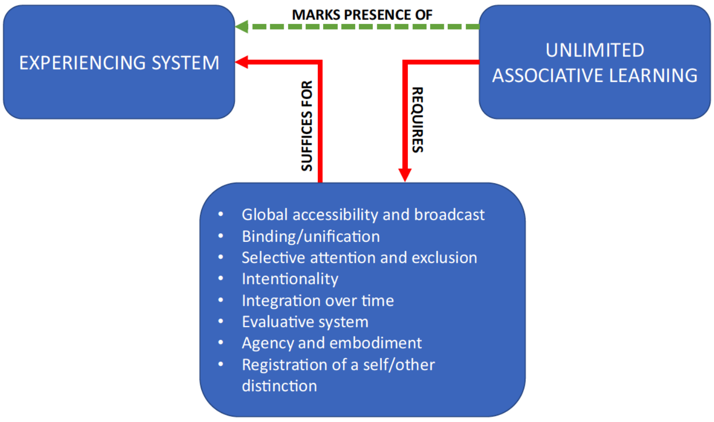

Figure 1. UAL as a transition marker for the evolutionary origin of consciousness (Figure 3 from Birch, Ginsburg and Jablonka 2020).

Crucially, Birch et al. adopt the above list of capacities as sufficient for consciousness on pragmatic grounds, that is, the list is not grounded in a particular theory of consciousness (say, global workspace theory). Instead, as noted, it represents a 'hidden consensus' that reflects various researcher's interests despite many of them disagreeing on the precise nature of consciousness. In other words, although there is little agreement regarding the best theory of consciousness, Birch et al. hold that these eight capacities are ones that "researchers would generally regard as *jointly sufficient* for a system being an experiencing subject" (Birch, Ginsburg and Jablonka 2020: 4).[1]

Further, they claim that UAL requires all eight of the capacities on the consensus list of conditions sufficient for consciousness (we are suspicious of this claim but will grant it for the sake of argument). Thus, UAL can serve as a marker or proxy for these capacities. Rather than investigating whether an organism has selective attention or can integrate information over time, one need only test whether that organism passes tasks requiring UAL. From this, Birch et al. emphasise that the five learning abilities comprising UAL form a 'natural cluster' (2020: 8). For example, we should not encounter an organism that can engage in trace conditioning

[1] By "sufficient" they mean "nomological sufficiency" or "*sufficiency in living organisms given the actual laws of nature*" (Birch, Ginsburg, and Jablonka 2020: 5, emphasis original).

but cannot do second-order conditioning. The abilities instead form a package or "a cluster of correlated abilities which are enabled by the overlapping underlying mechanisms" (Birch et al. 2020: 13).[2]

Before continuing, it is worth stepping back and highlighting the purpose of the UAL framework. The idea is that using UAL as a transition marker—and the set of capacities that come with it—can advance consciousness research without committing to a theory of consciousness. The framework purportedly does not make any commitments concerning the function of consciousness, how it works, its underlying mechanisms, etc. Instead, the two central commitments of the UAL framework (as presented by Birch, Ginsburg and Jablonka) are: 1) any organism exhibiting the eight hallmarks of consciousness should be viewed as conscious and 2) UAL requires the eight hallmarks of consciousness and thus can serve as a transition marker or proxy for identifying these capacities in organisms. In this way, the UAL framework is designed to be ecumenical. As they write, "agreeing on a transition marker is an important step for origins of consciousness research, because it allows researchers to unite around a shared agenda despite substantial disagreement about the nature of consciousness" (Birch, Ginsburg, and Jablonka 2020: 13).

Unfortunately, however, we do not think the story is so straightforward. As we argue below, we think the plausibility of the UAL framework *does* depend on making substantial theoretical commitments regarding the nature of consciousness. What we would like to suggest, then, is that without making such commitments—or, without being clear on which commitments one is making—then the scope of consciousness research the UAL framework promises to advance is unclear. Thus, while having a proxy for consciousness (both in an evolutionary setting and among currently existent organisms) does indeed expand the targets we might consider to be conscious, the scientific study thereof is typically seen as requiring not just descriptively or predictively adequate criteria, but also explanatorily justified reasons for thinking one organism (say, a bee) and not another (say, an *E. coli* bacterium) is conscious—and providing this leg of the story typically comes down to elucidating the architectural, mechanistic, and computational features of the system in question.

What is the upshot of this? In the rest of this paper, we argue for two separate but interrelated points: first, absent theoretical commitments to underlying structures or principles (such as we see in other theories of consciousness, and even earlier iterations of the UAL framework), the UAL framework is not sufficiently sensitive to important edge cases. That is, we might have reasons to believe organisms that exhibit some subset of the above eight capacities should be considered conscious, but the UAL framework remains silent when it comes to these

[2] The evidential status of this claim is somewhat unclear: Birch, Ginsburg, and Jablonka (2020) present it as a natural fact (8-9), an assumption of the UAL framework (13), as well as a hypothesis in need of empirical testing (13-14).

scenarios.[3] From this, we will argue that providing some underlying principle is important because it allows us to distinguish between *clutters* and *clusters* of properties: the former is a set of capacities or features unified by an underlying principle that ensures their unity; the latter, however, might be a heterogeneous ensemble of features only contingently related to each other (Mameli 2008). Determining this difference is a scientifically important step in consciousness science and requires the very theoretical commitments that Birch et al. wish to avoid, or so we argue. We will take each consideration in turn.

3. Sensitivity to Edge Cases

As introduced above, the UAL framework depends on what Birch, Ginsburg, and Jablonka pick out as a consensus list of eight sufficient conditions for consciousness. UAL serves as a marker for consciousness because all eight conditions are required for UAL. Presumably, then, any alternative or subset list of conditions recognized by other researchers is not recognised by the UAL framework. Are these eight capacities a good guide for identifying experiencing systems? There are reasons to think that they are not a good guide, at least not without additional theoretical commitments regarding how they are linked to consciousness.

An epistemic transition marker will face a familiar trade-off between sensitivity and specificity. Make the conditions too rigid, and you give up too much: "it is human" is a simple, reliable marker for conscious species, but leaves too much to the side. Conversely, focusing on associative learning (limited or otherwise) likely gets all the conscious organisms, but might include nonconscious ones such as plants and paramecia. Note that this is a trade-off specific to epistemic projects, as only they are attempting to use one property as a marker for another, and therefore must consider how informative each is about the other.

Birch and colleagues have picked out what seem like a fairly heavy-duty set of capacities. They thus err on the side of leaving out too much in order to have a clear picture. They emphasise that UAL is a positive marker and thus remains silent regarding which organisms lack consciousness. They also admit that their list of hallmarks may seem to some researchers to go "far beyond what is necessary" (Birch, Ginsburg, & Jablonka 2020: 19). UAL as a transition marker is thus not capable of adjudicating edge cases—cases in which different computational abilities dissociate from each other and from UAL. Instead, it only identifies those organisms that have the chosen list of eight hallmarks of consciousness.

Now, we submit that advocates of UAL as a marker of a major evolutionary transition *should* care about edge cases. For what is the null hypothesis that they ought to be concerned with? Presumably it is some sort of gradualist story. On a gradualist story, the eight capacities might have been cobbled together gradually and piecemeal among different organisms. One might still expect a set of useful capacities, and UAL too, to emerge at around the same time. That's

[3] For example, intentionality, agency, embodiment, and self-other registration. On some accounts (e.g., Merker 2007; Barron & Klein 2016; see also Irvine 2021) this subset is considered sufficient for consciousness—a point we turn to below.

not because these capacities are related to consciousness, or even to one another, but simply because organisms get complex enough to support them. An epistemic marker then merely picks a convenient point along what is properly seen as a gradual scale.

There are reasons to think that a gradualist story might be correct. Peter Godfrey-Smith, for example, advocates for such an account. He notes that gradualism is consistent with several accounts on the nature of consciousness. If consciousness is a form of information processing seen "from the inside", for example, then one would expect it to gradually evolve along with cognitive complexity (Godfrey-Smith 2017: 220; see also Godfrey-Smith forthcoming). Many other accounts of consciousness lend themselves to such as gradualist story: those that view consciousness as grounded in sensory-information processing, feelings or valuation, and information integration, for example (Godfrey-Smith 2017). Although we might expect some episodes of radical change to occur under these accounts (the evolution of the camera eye in the case of sensory-information processing, for instance), they are broadly consistent with a gradualist story.

There are also reasons to think that the eight sufficient conditions for consciousness identified by Birch et al. naturally come apart. For example, Elizabeth Irvine (2021) notes that a subset of the list of eight hallmarks (intentionality, agency, embodiment and self-other registration) is sufficient for consciousness under several models of consciousness, such as that advanced by Merker (2007) and Barron and Klein (2016). She writes that these four capacities "essentially turn out to refer to the capacity of an organism to generate egocentric representations of itself acting in space, where actions are goal-directed and selected in a top-down manner" (2021: 3). Compared to other potential markers for this package of capacities, Irvine notes that UAL is a poor choice. Godfrey-Smith also examines the idea that the sensory aspects of consciousness (e.g., point of view) and evaluative aspects of consciousness (e.g., pain) dissociate in some organisms. Some arthropods, like spiders and wasps, for example, might have rich sensory capacities while lacking equivalent evaluative experiences (Godfrey-Smith 2017). If these capacities do come apart, then they might represent different evolutionary routes to becoming an experiencing subject. As Godfrey-Smith writes,

> there are two traits here that have plausible connections to subjective experience, but they do not look like *different paths to the same thing*. They lead to different things. Both of them can be summarized with the idea that there is 'something it's like to be' one of those animals, but the evaluative and perceptual forms of this feeling-like-something are different. (Godfrey-Smith 2017: 224, emphasis original)

Godfrey-Smith notes that when it comes to gradualist views, "the aim is to leave behind simple 'in or out?' questions" and instead "to change how we think about that category" (forthcoming: 2).

Of course, the real test of these hypotheses would be to look to edge cases—that is, to organisms which have only some of the capacities, or partially unlimited learning—and see whether they are conscious. But, by picking a strong marker, Birch et al. have effectively ruled out this strategy: their marker remains silent precisely where we need it most.

To make the point another way, consider an analogy. Suppose comparative computer scientists of the far future are trying to discover when computers became Turing-complete. Evidence of actual architectures is dim and unreliable, but there are lots of screenshots from glossy magazines. A clever group of philosophers and scientists propose the following transition marker: computers are definitely Turing-complete once they have graphical user interfaces (GUIs). There's a good reason to believe this, too. A successful GUI usually requires a suite of capacities—sophisticated memory management, the capacity for multithreading, the ability to abstract using high-level programming languages—which are in turn pretty strongly associated with Turing completeness.

This is a good transition marker as far as it goes. Nevertheless, there are obvious gaps between the good epistemic condition this provides and the relatively poor metaphysical insight it gives on Turing-completeness. It is stronger than it needs to be since there are plenty of Turing-complete computers without GUIs. It also posits a cluster of features that arose together historically and are important for GUIs, but which can and do dissociate. Some of these features might be causally relevant to Turing-completeness, while others might be irrelevant. Now, all of these might be acceptable in some contexts. But if the dispute is with computational gradualists, they may well object. The gradualists claim that all computers worth their salt were Turing complete in some sense, though of course they vary in how much memory they have available. The gradual addition of memory made more of the Turing-computable functions actually computable, which in turn made the power of computers more obvious. Alongside this gradual addition came the addition of memory management and multithreading and the like—not because they are related to one another or to Turing-completeness, but simply because they are individually useful. Because each takes a certain amount of overhead, then chips have to reach a certain level of size and complexity before they can be implemented. But there are no major transitions in computation, only the relentless march of Moore's law. The gradualist should push back, as should anyone with a different picture of completeness.

4. Individually Sufficient versus Clutter

It is important to note that the above considerations about epistemic conditions hold even if there was only a single sufficient condition given for consciousness. However, the fact that there is a *set* of such conditions gives rise to an interesting dilemma. Take the set *S* of eight sufficient conditions—global broadcasting, intentionality, and so on. The question arises: do the members of *S* reliably co-occur or not?

Suppose first that they do: that is, that the presence of S_1 is good evidence for the presence of $S_{2,}$ and S_3 and so on. Reliable co-occurrence means that the presence of any member of the set is good evidence for the presence of the whole set. But then *S* is no longer a minimal set. Rather, there are a number of equally good minimal sets, each of which consists of a single member of *S.* In fact, any condition S^* that reliably correlates with any member of *S* will *also* constitute a minimally sufficient set. *S* is thus a jointly sufficient set of epistemic conditions for consciousness only in the trivial sense that it is a set of individually sufficient epistemic conditions. Hence, each individual condition can serve as a marker, and it is not clear why we need to rely on UAL.

Conversely, suppose that the members of S do not reliably co-occur. Then there is a point to enumerating all of them. But now it is unclear why it is *this* set and no other. The actual members appear to be something of a heterogeneous disjunction. To use a nice distinction from Mameli (2008), it raises the possibility that *S* is a mere *clutter* of conditions, rather than a real cluster of properties. Property clusters have a unity that is sustained by some underlying principle. Clutters don't.

To see why, consider first the Cluster Hypothesis. As Mameli remarks (2008: 720), defending the Cluster Hypothesis requires providing an account of an underlying principle that justifies the inclusion of the traits or properties into the same category, typically based on some kind of homeostatic property cluster *C*. A homeostatic property cluster *C* refers to a set of properties, capacities, or traits that co-occur as a result of an underlying causal process, which connect the properties in reliable ways. Thus, the Cluster Hypothesis is correct if one can enumerate a list of *i*-properties that constitute *C*: the causal processes "that connect such properties and cause them to tend to co-occur" (2008: 736). Determining the validity of the Cluster Hypothesis then turns crucially on the plausibility of the existence of the underlying homeostatic property cluster *C*. Conversely, if *C* does not exist, then we have reason to believe the *Clutter Hypothesis* is correct.

Translated into the case of UAL, we can see that unlimited associative learning itself cannot be the thing that provides unity: the arrow in Figure 1 goes from UAL to the set, not the other way around. Absent such a unifying principle, then—which epistemic conditions do not try to provide—we have at best a disjunctive set of jointly sufficient conditions. Determining the homeostatic property cluster *C*—and its constituent set of *i*-properties—is important for determining whether the set of eight sufficient conditions is a cluster or a clutter.

Thus, on either horn of the dilemma, we would appear to have a heterogeneous disjunction: on the first horn, a disjunctive set of individually sufficient conditions, on the second a disjunctive set of jointly sufficient conditions. This brings us to a meta-epistemic worry. Without some explanation about why these conditions *and not others*, our confidence that *S* is a good marker ought to be undermined. For heterogeneity is epistemically unsatisfying. At best, heterogeneous disjunctions suggest some kind of missed generalisation (Fodor 1997). We would be in a position of alchemists who had very many reliable tests for the presence of gold, but no idea whatsoever about which of them (if any) spoke to the nature of gold—and hence why any of them were decent tests.

Now, this may strike one as unfair to Birch et al. who offered *S* as a sort of working hypothesis. But the point is that even if *S* seems like it does a fair job on clear cases, we ought to have no confidence that it generalizes to *unclear* ones. Heterogeneity is an induction-breaker, absent some further story about the underlying mechanisms that support such an odd disjunction. The heterogeneity should also undermine our confidence that we have an epistemically sufficient set of conditions for identifying experiencing subjects.

5. An Objection and a Reply

Birch and colleagues might object by insisting that the eight hallmarks of consciousness are not heterogenous disjunctions. As noted in Section 2, they take the learning abilities that comprise UAL to form a natural cluster grounded in "overlapping underlying mechanisms" (Birch, Ginsburg and Jablonka 2020: 13). At several points they also suggest that this same set of mechanisms is responsible for consciousness. For example, they write, "the myriad of mechanisms underlying UAL in living organisms, constitute (are building blocks of and are nomologically sufficient for) biological consciousness" (10). Given that these mechanisms are "overlapping" or form a "package", then one might avoid the charge that the hallmarks of consciousness are a clutter. They are causally related in virtue of being grounded in a particular set of intertwined mechanisms—the same set that gives rise to UAL. Such a view also helps resolve edge cases. We can determine whether these mechanisms generate other behaviours—expanding S—or how they might come apart.

The idea that UAL and consciousness are grounded in the same set of mechanisms is supported by previous work. For example, Bronfman et al. (2016) provide a detailed functional architecture for UAL involving three interconnected mechanisms: feature-integration mechanisms, value systems and memory systems. They model how these mechanisms interact in the context of learning and consider how they might be implemented in neural structures and the body. It is because of these mechanistic considerations, that they suggest that the hallmarks of consciousness are best understood as arising from such a cluster of systems. They take UAL to provide a promising starting point for "reverse engineering" an experiencing system (Bronfman et al. 2016: 10). They also write, "We believe that our proposal provides an evolutionary-selective rationale for the emergence of the structures and processes suggested by prominent models of consciousness" (Bronfman et al. 2016: 10). Bronfman and colleagues go on to compare in detail their proposed architecture for UAL with contemporary theories of consciousness (such as global neural workspace, dynamic core theory, integrated information theory, and Merker's model of consciousness) (see also Bronfman et al. 2018, Ginsburg and Jablonka 2019). They show how their UAL model is compatible with these theories of consciousness.[4] In cases where their model is not compatible with a particular theory of consciousness, they suggest that this theory of consciousness might be mistaken (Bronfman et al. 2016: 11).

[4] Further, Ginsburg & Jablonka (2019) seem similarly committed to this match-up between UAL and theories of consciousness focused on architectures, computations, and mechanisms. For instance, they propose a hierarchical predictive processing [HPP] model of UAL because the former accommodates the phenomenological features of the latter (such as 'temporal thickness') explicitly within the way the HPP architecture itself is constructed. While the discussion of HPP is provisional, it is clearly an attempt at sketching the broad compatibility—and necessity—of some kind of enabling system to lend support to the UAL framework.

We think that the approach adopted by Bronfman, Ginsburg and Jablonka would avoid the concerns advanced in this paper (whether it works is a question for another day). But it would do so by providing a mechanistic story about the linkage between UAL and consciousness. And of course, in doing so, it makes substantial theoretical commitments regarding the nature of consciousness. In sum, then, it seems that the plausible ways around the objections we raise involve dropping back to metaphysics, which is precisely what Birch et al. wanted to avoid.

6. Conclusion

We have raised problems for Birch et al.'s search for epistemic transition markers. We suspect similar problems will plague any purely epistemic approach. But is that such a problem? On the surface, as we have remarked, there is much to recommend this approach because it plausibly expands the targets of investigation for consciousness research. Indeed, as noted above, Birch (2020) criticises what he calls 'theory-heavy' approaches to animal consciousness, which includes work by Merker (2007) and Barron & Klein (2016). One reason for Birch's scepticism is the *dilemma of demandingness* (2020: 6): stipulating too strong a sufficient condition (say, possession of an intact human neo-cortex) may certainly capture definitive instances of consciousness but provides no aid in the comparative study of animal consciousness. On the other horn of the dilemma, we might stipulate less stringent sufficient conditions (of which Merker's midbrain theory is an exemplar), but then the evidentiary link between this condition and consciousness is weakened "and the positive case for animal consciousness becomes correspondingly weaker" (ibid., see also Shevlin 2021). Theory-heavy approaches, then, are supposed to be of little assistance in advancing the study of animal consciousness.

To conclude, we will argue that the theory-heavy approach need not be as problematic as has been suggested. For one, consider the description of theory-heavy approaches: "We start with humans. We develop a well-confirmed, complete theory of consciousness in humans, and we take this theory 'off the shelf' and apply it to settle the question of whether animals, in disputed cases, are conscious or not" (Birch 2020: 2). Yet we suspect that no proponent of a theory-heavy approach would accept this characterisation. For one, even the most immodest theorist would struggle to assert that they have a 'well-confirmed, complete' theory of consciousness. Animal consciousness is interesting to think about in part because *human* consciousness is unsettled ground. Most of us hope that by studying humans we might shed light on animals, but also that by studying animals we might learn more about human consciousness.

In that sense, the comparative science of consciousness fits well with a picture of science on which identities are postulated as working hypotheses (Wimsatt 1974), a view McCauley & Bechtel term "heuristic identity theory". On such a view, nominally distinct scientific domains co-evolve by postulating and testing linkages between them. Importantly, these postulates are bi-directional: either science might revise its models, or the linkage itself might be denied, depending on the results of future experimentation and observation. Furthermore, any of the links in this chain might be tenuous and provisional. Indeed, such a position seems representative of earlier iterations of UAL. For instance, Ginsburg & Jablonka (2019) postulate a potential link between the features of UAL and hierarchical predictive processing models of cognition and consciousness (Friston 2018). They then sketch provisional linkages between the two, with the postulated model receiving some support from neurobiological models. While the predictive processing story might need revising, it is at least an attempt at furnishing a mapping between the epistemic transition marker of UAL and its enabling system.

Similarly, Barron & Klein (2016) take evidence that subcortical structures support consciousness in humans, combine this with Merker (2007)'s claims about the functional role of subcortical structures in supporting consciousness, and link that with recent work on the computational role of the central complex in insects to make a claim about invertebrate consciousness. Each of these links has evidence for it—but none of them could plausibly be characterized as 'well-confirmed' or 'complete.' By extending several plausible but speculative hypotheses to insects, their goal is to extend the set of systems that might be plausibly used to test the whole fabric of hypotheses—not to lean on obvious truths to establish surprising ones.

Indeed, we see no reason why a theory-heavy version of UAL might not also be pursued in the same vein. One might look to the list of capacities, or to UAL itself, to give a story about the *mechanism* of consciousness (as earlier version of UAL explore). As McCauley & Bechtel make clear, one can identify the mechanism for a phenomenon long before one can say anything about *why* that mechanism is sufficient. Making bold claims for UAL—bolder than an epistemic marker can support—is not the last step in a science of comparative consciousness but one of the first.

References

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 113(18), 4900–4908. https://doi.org/10.1073/pnas.1520084113

Birch, J. (2020a). In Search of the Origins of Consciousness: Simona Ginsburg and Eva Jablonka: The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness. MIT Press, Cambridge, MA, 2019, 646pp, ISBN: 9780262039307. *Acta Biotheoretica*, 68(2), 287–294. https://doi.org/10.1007/s10441-019-09363-x

Birch, J. (2020b). The search for invertebrate consciousness. *Noûs*, nous.12351. https://doi.org/10.1111/nous.12351

Birch, J., Ginsburg, S., & Jablonka, E. (2020). Unlimited Associative Learning and the origins of consciousness: a primer and some predictions. *Biology & Philosophy*, 35(6), 1-23.

Bronfman, Z. Z., Ginsburg, S., & Jablonka, E. (2016). The Transition to Minimal Consciousness through the Evolution of Associative Learning. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01954

Bronfman, Z. Z., Ginsburg, S., & Jablonka, E. (2018). The Evolutionary Origins of Consciousness. 28.

Crapse, T.B., & Sommer, M.A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9, 587-600.

Fodor, J (1997) Special Sciences: Still Autonomous after all these years.

Friston, K. (2018). Am I Self-Conscious? (Or Does Self-Organisation Entail Self-Consciousness?). *Frontiers in Psychology* 9:579. DOI: 10.3389/fpsyg.2018.00579.

Ginsburg, S., & Jablonka, E. (2015). The teleological transitions in evolution: a Gantian view. *Journal of Theoretical Biology*, 381, 55-60.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

Ginsburg, S., & Jablonka, E. (2021). Evolutionary transitions in learning and cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1821), 20190766. https://doi.org/10.1098/rstb.2019.0766

Godfrey-Smith, P. (2020). Varieties of subjectivity. *Philosophy of Science*, 87(5), 1150-1159.

Godfrey-Smith, P. (2017). The evolution of consciousness in phylogenetic context. In *The Routledge handbook of philosophy of animal minds* (pp. 216-226). Routledge.

Godfrey-Smith, P. (2021). Gradualism and the Evolution of Experience. *Philosophical Topics*.

Irvine, E. (2021). Assessing unlimited associative learning as a transition marker. *Biology & Philosophy*, 36:21, 1-5.

Jekely, G., Godfrey-Smith, P., & Keijzer, F. (2021). Reafference and the origin of the self in early nervous system evolution. *Philosophical Transactions of the Royal Society B*, 376: 20190764.

Mameli, M. (2008). On Innateness: The Clutter Hypothesis and the Cluster Hypothesis. *The Journal of Philosophy*, 105(12), 719-736.

McCauley, R.N. and Bechtel, W. (2001) Explanatory Pluralism and Heuristic Identity Theory. *Theory and Psychology* 11(6), 736-760.

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioural and Brain Sciences*, 30(1), 63-81.

Salmon, W. (1998). Causal and Theoretical Explanation. In *Causality and Explanation*. New York: Oxford University Press, pp108-124.

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. Mind & Language, 36(2), 297-314.

Smith, J.M. and Szathmáry, E. (1995) *The major transitions in evolution.* New York: Oxford University Press.

Wimsatt, W. (1974) Reductive explanation: A functional account. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1974: 671-710.