Synthese (2024) 203:171
https://doi.org/10.1007/s11229-024-04539-6

ORIGINAL RESEARCH

A credence-based theory-heavy approach to non-human consciousness

C. R. de Weerd[1,2]

Received: 3 September 2023 / Accepted: 20 February 2024
© The Author(s) 2024

Abstract
Many different methodological approaches have been proposed to infer the presence of consciousness in non-human systems. In this paper, a version of the *theory-heavy* approach is defended. Theory-heavy approaches rely heavily on considerations from theories of consciousness to make inferences about non-human consciousness. Recently, the theory-heavy approach has been critiqued in the form of Birch's (Noûs 56(1):133–153, 2022) dilemma of demandingness and Shevlin's (Mind Lang 36(2):297–314, 2021) specificity problem. However, both challenges implicitly assume an inapt characterization of the theory-heavy approach. I argue that an alternative characterization of the approach, what I call a *credence-based theory-heavy* approach, avoids these challenges. Theorists can generate interpretations of their theory, at different levels of generality, and operationalize these into theory-informed markers. These theory-informed markers are assigned a likelihood and are used to assess the probability that a target system is conscious. In providing this characterization, and mapping out the possible ways in which a credence-based theory-heavy approach can be fleshed out, the aim is to situate the theory-heavy approach as a more compelling approach than it is currently being perceived as. Our attention, then, needs to shift towards remaining challenges such as the consensus problem and the problem of calibrating the likelihoods associated with theory-informed markers. I also explore methodological pluralism and assess how the credence-based theory-heavy approach can benefit from other methodological approaches.

Keywords Phenomenal consciousness · Non-human consciousness · Animal consciousness · Measuring consciousness · Theory-heavy approach

✉ C. R. de Weerd
christian.r.de.weerd@gmail.com

[1] Department of Theoretical Philosophy, University of Groningen (RUG), Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

[2] Centre for Philosophy and AI Research (PAIR), University of Erlangen-Nürnberg (FAU), Werner-von-Siemens-Str. 61, 91052 Erlangen, Germany

Published online: 13 May 2024

1 Introduction

The focus of this paper is on the so-called distribution question (Allen & Bekoff, 1997, 2007; Prinz, 2005): Which animals (or other non-human systems) are conscious?[1] *Prima facie,* there is a strong intuition that animals like chimpanzees are conscious. But what justifications do we have, other than pre-theoretical intuitions, to ascribe consciousness to such animals? What about more controversial cases in which our intuitions are not as strong, for instance in the case of a garden snail (Schwitzgebel, 2020)? As it stands, there is no consensus on which methodological approach is best suited to answer these questions. Despite this, a wide range of methodological approaches are available. Before I describe which methodological approach I will defend, consider first the following brief overview of the methodological landscape. This overview is meant to give a general impression of what kind of methodological approaches are available. Throughout the paper, approaches will be described in more detail whenever necessary.

Birch (2022) identifies the following three methodological approaches: a *theory-heavy*, *theory-neutral*, and *theory-light* approach. Birch characterizes the theory-heavy approach as proposing that "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, 2022, p. 134).[2] On the other hand, a theory-neutral approach "build[s] up a list of behavioural, 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, 2022, p. 134).[3] A theory-light approach, importantly endorsed by Birch himself, "commit[s] to a broad hypothesis about the relation between phenomenal consciousness and cognition that is compatible with a wide range of more specific theories" (Birch, 2022, p. 140). A fourth methodological approach that is not often explicitly recognized can be referred to as an *intuition-based* approach. The main idea is that pre-theoretical intuitions must, in some way, guide our methodological approach to animal consciousness. Guiding can be understood in a *direct* or *indirect* sense. Sometimes pre-theoretical intuitions are used directly to infer consciousness in non-human systems.[4] However, more often pre-theoretical intuitions are used indirectly to constrain or dismiss theories/criteria that are either very liberal (e.g., Tononi & Koch, 2015) or very conservative (e.g., Carruthers, 1999)

[1] Within the debate, it is conventional to conceptually distinguish *phenomenal* consciousness from *access* consciousness (Block, 1995). A state is access conscious if it's available for further reasoning, or guiding verbal report and action. A state is phenomenally conscious if there is something it is like to be in that state (Nagel, 1974). The distribution question is normally taken to concern the distribution of phenomenal consciousness. Therefore, unless specified otherwise, the term consciousness will refer to phenomenal consciousness.

[2] Some examples of exploring a theory's implications on non-human systems include Seacord (2011), Carruthers (1999), Gennaro (2004) and Tononi and Koch (2015). Also, see Dennett (1995) for an explicit endorsement of this approach.

[3] A prime example of this approach is Tye's (2016a, 2016b) use of *Newton's Principle* to infer animal consciousness.

[4] Such an approach can be grounded in so-called perceptualism (e.g., Jamieson, 1998); the idea that properties of a system's mental state are contents of our perception of those systems (Allen, 2016).

in their attribution of consciousness to non-human systems (Schwitzgebel, 2020).[5] The aforementioned methodological approaches need not necessarily exclude one another. In fact, it will arguably be fruitful for different methodological approaches to work together. Let *methodological pluralism*, then, refers to the usage of a variety of methodological approaches in our search for non-human consciousness.[6] Shevlin (2021) has articulated a methodological proposal along these lines which he coined the *modest theoretical* approach in which "theory-heavy and theory-light approaches can operate in a form of dynamic equilibrium, with the insights of each informing and constraining the other" (Shevlin, 2021, pp. 308–309).[7]

Recently, two challenges in the form of the *dilemma of demandingness* (DOD) (Birch, 2022) and the *specificity problem* (Shevlin, 2021) have been raised against the theory-heavy approach. Both challenges are prevalently used against the theory-heavy approach, but implicitly rely on an arguably inapt characterization of the theory-heavy approach. The goal of this paper is to provide a characterization, or reinterpretation, of the theory-heavy approach that avoids these challenges. I call this the *credence-based theory-heavy* approach. Moreover, another aim of this paper is to carve out the space of possible ways a credence-based theory-heavy approach can be fleshed out, thereby encouraging theorists to think about how to draw implications from their theory within a credence-based context.

The paper is structured as follows. In Sect. 2, I describe how the DOD and the specificity problem pose problems for the ordinary theory-heavy approach. In Sect. 3, I propose the credence-based theory-heavy approach, which combines a probabilistic stance towards inferences about consciousness, and a pluralistic stance towards interpretations of a theory. Subsequently, I use this credence-based theory-heavy approach to answer the DOD and the specificity problem. In Sect. 4, I discuss some remaining challenges for the credence-based theory-heavy approach and assess the possibility of *methodological pluralism* in which the credence-based theory-heavy approach forms a coalition with other methodologies. Section 5 concludes.

2 Two problems for the theory-heavy approach

Typically, theories of consciousness begin by describing the structure of *human* consciousness and how it relates to the human brain. Such theories are usually motivated, and validated, by empirical findings in humans and are mostly based on verbal reports about introspective processes.[8] Only after that, predictions are made about non-human systems based on theoretical considerations. This process is often referred to as a theory-heavy approach to consciousness (Birch, 2022). A prime example is the claim by Tononi & Koch that "the more the postulates of IIT are validated in situations in which we are reasonably confident about whether and how consciousness changes, the

[5] Some prime examples include Aaronson's (2014) intuition-based argument against IIT and Shevlin's (2021) intuition-based rejection of a liberal and conservative characterization of the theory-heavy approach.

[6] I will discuss methodological pluralism in more detail in Sects. 4.1 and 4.3.

[7] Recently, Dung (2022) has produced a version of the modest theoretical approach in the form of a list of eight desiderata used to assess tests of animal consciousness.

[8] Although, see Sect. 4.2.2 for a brief discussion on the use of no-report paradigms.

more we can use the theory to extrapolate and make inferences about situations where we are less confident" (Tononi & Koch, 2015, p. 10).[9] However, this human-centered approach (Veit, 2022a) has been criticized for being too human-centered, and it has been suggested that "without the help of a snail's introspection or verbal reports, it is unclear how we should then generalize such findings to the case of the garden snail" (Schwitzgebel, 2020, p. 58). The underlying reasons for this supposed unsuitability of a theory-heavy approach to make extrapolations beyond the human case are helpfully illustrated by the following two challenges. Birch (2022) poses a so-called:

> **Dilemma of demandingness (DOD):** "Strong sufficient conditions will not get us very far in making inferences about cases other than humans who can report their experiences, if they get us anywhere. Yet as we formulate increasingly weaker conditions, the evidence from humans that they amount to a sufficient condition become increasingly weaker, and the positive case for animal consciousness becomes correspondingly weaker" (Birch, 2022, p. 138).

According to Birch, theories that have a strong human evidentiary basis specify sufficient conditions so strong that they cannot be applied to non-human systems. For instance, global workspace theory (GWT) posits that human consciousness is explained by a global broadcast mechanism that integrates information from perceptual, affective, and memory systems and broadcasts this information back to the input and consumer systems (including verbal report, planning, reasoning, and decision making) (Dehaene & Changeux, 2011, p. 209). However, Birch (2022, p. 136) argues that such a cautious interpretation of the theory "remains silent about cases in which something less than the full network is present", since it remains unclear, due to a human evidentiary basis, which consumer systems can be taken offline for a global broadcast mechanism to still elicit consciousness. Consequently, theories with strong sufficient conditions, but also strong human evidentiary basis, cannot provide much insight into non-human cases. Theories specifying weaker sufficient conditions [e.g., mid-brain theory (Merker, 2007)] face the opposite problem. They might be applicable beyond the human case but lack evidentiary support. For instance, empirical support for the mid-brain theory (Merker, 2007) may depend on evidence from humans without a functional cortex, which is much harder to come by (Birch, 2022, p. 137). Absent empirical confirmation, it is unclear whether predictions about non-human systems made by such theories are reliable or not. According to Birch, then, theory-heavy proponents inevitably have to make "a trade-off between the relevance of these conditions to animals, and the strength of the evidence for their sufficiency" (Birch, 2022, p. 145).

Shevlin (2021) poses a similar[10] challenge that he calls the:

> **Specificity problem (SP):** "The specificity problem is the challenge of how to spell out the cognitive mechanisms identified as constitutive of consciousness

[9] See also Seacord (2011).

[10] Importantly, Shevlin's challenge is distinct in focus from Birch's. Where Birch focuses on the demanding nature of sufficient conditions, Shevlin focuses on the lack of principled ground to locate *the* appropriate level of specificity of the proposed cognitive mechanism constitutive for consciousness.

according to our preferred theory in such a way as to make them applicable beyond the human case" (Shevlin, 2021, p. 300).[11]

Shevlin argues that it is not clear at *which specific* level of specificity, or abstraction, the relevant cognitive mechanism must be specified. For instance, GWT proponent Dehaene claims that "consciousness is just brain-wide information sharing" but also that "the capacity to report is a key feature of a conscious state" (Dehaene, 2014, p. 165). The former is a very abstract, or coarse-grained, description of the theory whilst the latter specifies a much more detailed, or fine-grained, feature. However, our human evidentiary basis is compatible with both interpretations and thus *underdetermines* what the right level of specificity is. Thus, the following open question arises: Is mere brain-wide information sharing enough to elicit consciousness, or does the broadcast network require, for instance, the full set of human consumer systems? If this question remains unanswered, Shevlin argues, it is unclear whether non-human systems that employ, for instance, a more coarse-grained version of the respective cognitive mechanism are conscious (Shevlin, 2021, p. 301). Crucially, then, the SP is only problematic if you accept that the right level of specificity *must* be identified first before any predictions about non-human consciousness can be made. This assumption, as well as Birch's assumption that theory-heavy proponents need to produce sufficient conditions, will be questioned in the following section.

3 A credence-based theory-heavy approach

3.1 An overly demanding characterization of the theory-heavy approach

Birch and Shevlin raise important challenges for the theory-heavy approach. However, they target a specific characterization of the theory-heavy approach that is, admittedly, too demanding and implausible. Recall that Birch characterizes the theory-heavy approach as follows: "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* [emphasis added] the question of whether animals in disputed cases, are conscious or not" (Birch, 2022, p. 135). Elsewhere he wonders, "*exactly* [emphasis added] how similar to the human global workspace does a workspace have to be to *suffice* [emphasis added] for conscious experience?" (Browning & Birch, 2022, p. 4). But this, alongside the rest of Birch's discussion on the DOD, implies that theorists are committed to formulating a set of (stringent) sufficient conditions and applying these in a binary, definitive, fashion to animals; based on these exact sufficient conditions the target system is either conscious or not. Similarly, Shevlin argues that making theory-based predictions about non-human consciousness "would only be possible, however, once we had some grounds for spelling out GWT at an appropriate level of specificity" (Shevlin, 2021, p. 310). This implies that theories can only determine, in

[11] As Shevlin also observes, the specificity problem has been recognized by others (e.g., Carruthers, 2018a, 2018b; Prinz, 2018; Schwitzgebel, 2020), but Shevlin is the first to propose this as a *general* problem for the theory-heavy approach.

binary or definitive fashion, which non-human systems possess consciousness once it is clarified what the appropriate level of specificity is.

These considerations reveal that Birch and Shevlin characterize the theory-heavy approach as adopting a *binary stance* towards inferences about non-human consciousness, and a *singular stance* towards sufficient conditions or levels of specificity. Adopting a binary stance means that only definitive statements can be made about consciousness; the target system is either conscious or not. Adopting a singular stance means that one set of sufficient conditions, or one level of specificity, must be identified. However, this combination commits proponents of the theory-heavy approach to specify *exactly* what their theory's sufficient conditions are, or level of specificity is, after which they must use this to make binary, definitive, predictions about non-human consciousness. But succeeding in such a task is too demanding for *any* approach to non-human consciousness.

3.2 Towards a credence-based theory-heavy approach

Fortunately, a more modest characterization of the theory-heavy approach is available, which I refer to as the *credence-based theory-heavy* approach. Concisely, the credence-based theory-heavy approach entails that theorists should (1) adopt a *probabilistic*, instead of a binary, stance towards inferences about non-human systems. A probabilistic stance entails making inferences about non-human consciousness in terms of likelihoods, or quantified degrees of confidence. Moreover, theorists should (2) adopt a *pluralistic*, instead of a singular, stance towards different interpretations of their theory. A pluralistic stance entails utilizing multiple interpretations of a theory, differing in degree of generality (or, similarly, level of specificity), to infer consciousness in non-human systems. Before spelling out what it will look like for a theory-heavy proponent to adopt both stances, consider briefly the following motivations for adopting these stances in the first place.

Many alternative approaches to non-human consciousness already take some kind of probabilistic stance. For instance, Dung and Newen (2023) have proposed a two-tier framework in which the distribution of consciousness is inferred with weak and strong indicators, expressing the likelihood that the target system is conscious. Moreover, the quality of conscious experience is categorized via a set of dimensions which are assigned a numerical score based on "the extent to which different operationalizations suggest that animals possess features relevant to consciousness" (Dung & Newen, 2023, p. 8).[12] All scored dimensions collectively generate a comprehensive consciousness profile. Greater scores correspond to a heightened level of confidence regarding the presence of (the specific dimension of) consciousness in the target system. In a similar probabilistic spirit, Dung (2022) has proposed a list of eight desiderata that can be used to assess the strength of animal consciousness tests. The more desiderata that are satisfied by a test, the more confidence we can have in its predictions about the presence of consciousness in target systems. Lastly, Birch's version of his theory-light approach involves a so-called *facilitation hypothesis* according to which phenomenal

[12] See Birch et al. (2020) for a similar approach.

consciousness likely facilitates a cluster of cognitive capacities such as trace conditioning, reversal learning, and cross-modal learning (Birch, 2022, p. 140). Importantly, "the larger the fraction of the cluster we find in a given species, *the stronger the case* [emphasis added] for consciousness will be in that species" (Birch, 2022, p. 145). Taken together, these approaches do not commit themselves to making binary, definitive, predictions, but rather adopt a probabilistic stance. As I will soon demonstrate, adopting a probabilistic stance is also something that a theory-heavy proponent can do.

The pluralistic stance partly finds its inspiration in discussions on *models of consciousness*. Here, it is argued that full-fledged, or fine-grained, models of consciousness might apply to humans, but provide a poor model for consciousness in non-human systems (Wiese, 2020, 2023). This is because such fine-grained models contain details based on human architecture, or human-specific properties of consciousness, that do not apply to non-human systems. In addition, it is likely that many of these details do not play a constitutive role in instantiating, or make a difference to, consciousness (Klein et al., 2020). A push towards generality is therefore encouraged in which models abstract away from, for instance, human-specific properties of consciousness. Subsequently, these models can be applied to a wider range of systems. On a general level, the relevant insight here is that more abstract, or coarse-grained, models of consciousness should also play *some* role in inferring consciousness in non-human systems. The pluralistic stance embodies this by attributing a role to multiple interpretations of a theory, that differ in their level of grain/abstraction, to infer consciousness in non-human systems.[13]

3.3 Explicating the credence-based theory-heavy approach

It is now possible to articulate the credence-based theory-heavy approach in more detail and show how a theory-heavy approach can adopt a probabilistic, as well as a pluralistic, stance. First, theorists can adopt a pluralistic stance in the following way. They are not committed to specifying only one level of their candidate mechanism for consciousness, or one set of sufficient conditions. Rather, they can generate a wide range of interpretations of their theory, or theory-based models, ranging from full-fledged (or fine-grained), to very abstract (or coarse-grained). Each interpretation, or model, can subsequently be operationalized into a *theory-informed marker*. In essence, theory-informed markers are features specified by a model that can be found in a target system. These theory-informed markers can subsequently be used to make inferences about consciousness in non-human systems; the presence of a theory-informed marker in a target system can raise the probability that a target system is conscious. Importantly, theory-informed markers can be operationalizations of the relevant cognitive mechanism at a certain degree of generality [e.g., brain-wide information sharing (Dehaene, 2014)]. However, theory-informed markers can also

[13] Notice, however, that the pluralistic stance differs crucially from finding *a* model that finds the right balance between being general, or coarse-grained, yet informative (Wiese, 2023). Instead, the pluralistic stance recognizes that *many different* models, differing in their level of grain, can be used simultaneously to make inferences about consciousness in non-human systems.

operationalize necessary conditions for consciousness as specified by the theory [e.g., internal consistency of globally broadcasted messages (Baars, 1988)]. The presence of a theory-informed marker that operationalizes a necessary condition *alone* (absent a theory-informed marker that operationalizes the relevant cognitive mechanism) only provides a very weak indication that the target system is conscious. It may well be that both *kinds* of theory-informed markers need to be present to have any *strong indicators* at all, depending on how much stock the theorist puts in these necessary conditions.[14] Taken together, theorists should generate multiple interpretations or models of their theory, at various levels of generality, develop associated theory-informed markers, and use these markers to infer the presence of consciousness in target systems.

However, not every theory-informed marker is an equally strong indicator of consciousness. Recall that theories are usually motivated, or validated, by empirical findings in humans and typically start with describing the structure of human consciousness, and how it relates to the human brain. Thus, the presence of a full-fledged theory-informed marker that encompasses all human-specific properties of consciousness is typically a very strong indicator of consciousness. But the presence of a coarse-grained theory-informed marker that merely operationalizes a very abstract, or liberal, interpretation of the theory is a weak indicator of consciousness since it is more *likely* that relevant details are abstracted away from (see Klein et al., 2020). Despite this, such a weak indicator still gives us *some* reason to think that the target system might be conscious. To capture these differences between markers a probabilistic stance must be adopted. Theory-informed markers that more closely resemble a full-fledged theory-informed marker are typically assigned a higher likelihood than more abstract theory-informed markers.

These *weighted* theory-informed markers can subsequently be used to develop a so-called:

> **Theory-based probability space:** Theories generate a probability space in which the likelihood that target systems are conscious is depicted. These likelihoods are determined by the presence of weighted theory-informed markers.[15]

Each theory of consciousness can generate its own theory-based probability space. A theory-based probability space simply consists of the following two elements (Fig. 1, Appendix). First, the x-as describes a set of target systems (e.g., chimpanzees, dogs, snails, etc.) that a theorist wants to investigate and compare. Second, the y-as describes how likely it is that the target system is conscious, based on the presence (or absence)

[14] This depends on how convinced the theorist is that their specified necessary conditions *must* be met. Suppose a theorist is very convinced that necessary condition X, operationalized in a theory-informed marker (TIM-X), must be met. Suppose further that this theorist generates a very fine-grained interpretation of their theory Y that operationalizes in a theory-informed marker (TIM-Y). Then, only the co-appearance of TIM-X and TIM-Y suffices as a strong indicator of consciousness in a target system. If a system only exhibits TIM-Y, but not TIM-X, this theorist can argue that the mere presence of TIM-Y is a weak indicator of consciousness since TIM-X is missing, *even though TIM-Y is a fine-grained theory-informed marker*. See p. 10 for an example. In any case, it is expected that the more fine-grained the exhibited theory-informed markers are, the more likely it is that these necessary conditions are also met.

[15] See Fig. 1 in Appendix for a visualization of what a theory-based probability space could look like. The proposal is general, but in principle any theory and any species or target system can be injected into the model.

of theory-informed markers. The best way to quantize likelihoods is up for debate and, at least to some extent, depends on context and methodological preferences. The choice here will be a "pragmatically motivated idealization" (Dung & Newen, 2023, p. 2). For instance, it is possible to use probabilities (e.g., between 0 and 1), a numerical scoring system (e.g., 1–5), or categories (e.g., weak, moderate, and strong indicators). It is also possible to avoid assigning likelihoods to systems directly, and instead create an ordered set that ranks target systems from most to least likely to be conscious. In any case, how the likelihood of a theory-informed marker is to be determined "cannot be fully captured by an algorithmic procedure. For it depends on how the different operationalizations should be weighted which in turn depends on context factors, in particular the specific species that is examined, interdependencies between different operationalizations and the reliability of the particular set of studies under scrutiny. Thus, such judgments should eventually be left to subject matter experts" (Dung & Newen, 2023, p. 8). Having said that, the general principle holds that more fine-grained theory-informed markers should typically be assigned a higher likelihood than more coarse-grained theory-informed markers. Moreover, in Sect. 4 I will discuss, on a more general level, how these likelihoods may be calibrated.

Consider the following simplified example to illustrate how this approach could work in practice. Suppose that (you are convinced that) GWT is true.[16] A credence-based theory-heavy approach to non-human consciousness using GWT, then, would work roughly in the following way. Recall that at least two interpretations of GWT are available. According to one interpretation, a full-fledged human global broadcast mechanism is required that integrates information from perceptual, affective, and memory systems and broadcasts this information back to the input and consumer systems (including verbal report, planning, reasoning, and decision making) (Dehaene & Changeux, 2011, p. 209). The associated theory-informed marker (TIM-A), then, would be the presence of a complex information sharing structure that contains all these systems. Suppose, then, that system W contains such a structure. Since this theory-informed marker resembles a full-fledged human global broadcast mechanism so closely, it is assigned a high likelihood (e.g., 0.9 or *very likely*) that system W is conscious. On another interpretation of GWT, "consciousness is just brain-wide information sharing" (Dehaene, 2014, p. 165). The associated theory-informed marker (TIM-B), then, would be any simple neuronal structure that can transmit and share information. Suppose, then, that system X contains such a structure (but lacks further fine-grained details). Since the interpretation is very abstract and coarse-grained, it is assigned a low likelihood (e.g., 0.2 or *not very likely*) that system X is conscious. Despite being a weak indicator, the presence of such a structure gives us at least *some* reason to think that such a system is more likely to be conscious than a system lacking any global broadcast mechanism whatsoever. It is also possible to generate a theory-informed marker (TIM-C) based on a necessary condition that GWT specifies, for instance that globally broadcasted messages must be internally consistent (Baars, 1988). Suppose, then, that system Y only meets this condition and lacks TIM-A or

[16] It is, of course, highly controversial whether or not GWT is true. See Sect. 4.2.1 for a discussion on this issue. Moreover, it can still be useful for those not convinced by GWT, or any other theory, to use the credence-based theory-heavy approach to extract implications from a theory for non-human systems. For instance, to show that the theory has counterintuitive implications (see Sect. 4.1).

TIM-B. In that case, the presence of such internal consistency *alone* would be a very weak indicator that the system is conscious. However, it is possible that a GWT theorist strongly believes that TIM-C must be satisfied for a system to be conscious. Now suppose that system W exhibits both TIM-A and TIM-C. In that case, the co-appearance of both markers constitutes a strong indication that system W is conscious. However, suppose that system Z exhibits TIM-A but not TIM-C. In that case, the absence of TIM-C and the sole presence of TIM-A might only constitute a weak reason to think that system Z is conscious.

Our discussion of the theory-informed markers so far has primarily been on how our credence should change if a theory-informed marker is present. What about negative evidence? Should the absence of theory-informed markers also move our credence on this approach?[17] This depends on what a theorist's stance is on what we might call the:

> **Asymmetry thesis:** The presence of an indicator is capable of moving our credence that a target system is conscious, whereas the absence of an indicator is not.

Many adopt an asymmetry thesis of some kind. Consider, for instance, the debate on fish pain. Key (2016) has argued that fish do not experience pain on the basis of lacking neural structures taken to be relevant for pain in humans. A widespread response is to argue that fish might implement pain via different neural mechanisms, and that the absence of neural structures which are relevant for pain in humans does not rule out this possibility (e.g., Elwood, 2016; Godfrey-Smith, 2016; Manzotti, 2016; Seth, 2016; Striedter, 2016). Thus, the argument goes, the absence of the aforementioned neural structures should not move our credence that the system is conscious. Underlying this response is the idea that "very different neuronal systems can serve a common function". Thus, merely observing the lack of neural structures alone cannot suffice to rule out that fish have consciousness. Theories of consciousness, however, can do more than that. Theorists can spell out what the function of consciousness is, articulate the necessary conditions to fulfill this function, and transform these into theory-informed negative markers that specify which kind of structures need to be present to implement this function. Theorists then, at least *prima facie*, have a principled way to develop negative markers whereas this is arguably more problematic for the theory-light and theory-neutral approach (see Andrews, 2024). However, developing negative markers will be a harder task for any theory of consciousness than developing positive markers (Birch, 2022). Moreover, on one way of interpreting Michel's (2019a) criterion problem it also remains unclear whether necessary conditions specified by theories of *human* consciousness are also necessary conditions for consciousness in non-human systems, since it could be that consciousness serves a different function in animals. In that case, failing to satisfy negative markers derived from a theory should not move our credence that a target system is conscious.[18] Whether the absence of a theory-informed

[17] I want to thank an anonymous reviewer for posing this question.

[18] There is a further potential complication here that it can be that certain necessary conditions spelled out by a theory are only necessary for certain *dimensions* of *human* consciousness. A system failing to satisfy this condition, then, might lack this dimension (e.g., unity) but might still have a more primitive form of consciousness (e.g., valenced consciousness) (Veit, 2022b).

marker should move our credence, then, depends on what stance the theorist takes in this debate, and whether or not they think that the asymmetry thesis also applies to theory-informed indicators.

To recapitulate, I have proposed a credence-based theory-heavy approach that incorporates a pluralistic stance and a probabilistic stance; theorists should generate multiple interpretations of their theory, develop associated *weighted* theory-informed markers, and use these markers to infer the presence of consciousness in target systems.[19] The outcomes can subsequently be depicted in a theory-based probability space. This characterization of the theory-heavy approach is not too demanding and takes into consideration that "even the most immodest theorist would struggle to assert that they have a well-confirmed, complete, theory of consciousness" (Halina et al., 2022, p. 75). In the following section, I will show how the credence-based theory-heavy approach deals with Birch's DOD and Shevlin's specificity problem.

3.4 Revisiting the dilemma of demandingness and the specificity problem

The credence-based theory-heavy approach answers both Birch's and Shevlin's challenges. Recall that the underlying cause for the DOD is that according to Birch theorists are committed to specifying sufficient conditions: "However, possession of a full human global broadcast network is a cognitively demanding sufficient condition that no non-human animal can meet, and the GWT does not tell us how much we can weaken these demands *and still have a sufficient condition* [emphasis added]. The result is that the theory *cannot settle* [emphasis added] disputes about animal consciousness" (Birch, 2022, p. 138). However, the credence-based theory-heavy approach adopts a pluralistic and probabilistic stance, and is thereby not committed to producing sufficient conditions, nor does it need to make definitive judgments. Instead, theorists focus on developing a wide range of weighted theory-informed markers. These theory-informed markers are much less demanding than sufficient conditions since theory-informed markers need not *settle* disputes about animal consciousness but can simply help to make a probabilistic, or credence-based, inference. To put it succinctly, Birch's theory-light approach "avoids the dilemma by avoiding altogether the attempt to construct sufficient conditions for consciousness" (Birch, 2022, p. 145), and I have shown that such a move is also available for the theory-heavy approach.

[19] It is appropriate to point out that during the late-stage development of this paper, Butlin et al. (2023) released a pre-print wherein they independently suggest a similar approach in which *indicator properties* of consciousness are derived from theories to assess whether AI systems are conscious. It is worth briefly pointing out some differences. In addition, *indicator properties* need to be described in computational terms, whereas it is also possible to describe *theory-informed markers* in non-computational terms. Moreover, the way in which *theory-informed markers* are derived from theory is different and arguably more principled. *Indicator properties* are derived by directly looking at what theories or theorists claim is necessary or sufficient for consciousness. However, *theory-informed markers* are derived via a two-step process. First, the mechanism described by the theory is articulated, or interpreted, at different levels of granularity. Only afterwards are these interpretations operationalized into theory-informed markers. Lastly, Butlin and colleagues bundle indicator properties from different theories together to make inferences about AI systems. However, the credence-based theory-heavy approach, at least in its standard formulation, focusses on extracting predictions from one theory at the time. See Sect. 4.2.1 for a deeper discussion on how the credence-based theory-heavy approach can accommodate insights from different theories.

The credence-based theory-heavy approach also deals quite naturally with the specificity problem. To understand why, consider first Shevlin's discussion of two potential solutions a theory-heavy proponent can appeal to.[20] First, a theory-heavy proponent might give a *conservatist response.* Assuming no evidence to the contrary, "the full range of capacities associated with our candidate mechanism for consciousness in humans are essential" (Shevlin, 2021, p. 301).[21] For instance, a GWT proponent could claim that all human capacities associated with globally broadcast information (e.g., reflective thought, deliberation, self-awareness, verbal report) are crucial for consciousness. However, Shevlin dismisses this response because it produces counterintuitive predictions (i.e., excluding systems that intuitively ought not to be excluded) and it seems under-motivated (Shevlin, 2021, p. 302). Second, a theory-heavy proponent might give a *liberal response.* Here, it is claimed that "consciousness [is identified] with the cognitive capacities and mechanism proposed by the theory in question at the greatest possible degree of generality" (Shevlin, 2021, p. 302).[22] For instance, a GWT proponent could claim that any form of system-wide information sharing elicits consciousness. Shevlin dismisses this response mainly based on its counterintuitive consequences (i.e., including systems that intuitively ought not to be included) (Shevlin, 2021, p. 303).

Shevlin's appeal to counterintuitive consequences *prima facie* warrants initial skepticism towards both the conservatist and liberalist approaches, and for some might even be a sufficient reason to reject them.[23] However, there is a better, more principled reason, to reject both the conservatist and liberal response. Namely, that neither the conservative nor the liberal response is *necessitated*, or *entailed,* by their preferred theory of consciousness. Notice how both the conservative and liberal responses solve the SP by simply committing to their preferred level of specificity. But the SP reveals that we do not know which level of specificity is the correct one. Thus, both the conservative and liberal responses simply beg the question against SP by just picking a level of specificity. The problem, then, fundamentally lies in the fact that both the conservative and liberal approach adopt a singular stance, and thereby commit themselves to one level of specificity. However, the pluralistic stance, adopted by the credence-based theory-heavy approach, solves this problem by not having to commit to *one* level of specificity. Instead, it attributes a role to different levels of specificity in making inferences about non-human consciousness. Therefore, there is no need to identify the right level of specificity first, and, as such, it does not force us to be either too liberal or too conservative. In fact, it incorporates both responses in the following way. Theory-informed markers, developed from liberal interpretations of a theory, are

[20] Shevlin (2021) also discusses two other responses a theory-heavy proponent might give, namely an *incrementalist* and *rejectionist* response (see pp. 303–305). These responses, along with their refutations, however, are not relevant for the present purposes.

[21] To see how this response contrasts with the credence-based theory-heavy approach, see Fig. 2 in Appendix.

[22] To see how this response contrasts with the credence-based theory-heavy approach, see Fig. 3 in Appendix.

[23] However, others argue that intuitions should play no role (e.g., Murray, 2020). Moreover, see Sect. 4.1 for a discussion on why intuitions can be misleading since it's not always clear what intuitions target.

assigned a low likelihood whereas theory-informed markers developed from conservative interpretations of a theory are assigned a high likelihood. Making inferences about consciousness in non-human systems does *not* require us to commit to one level of specificity, contrary to what Shevlin (2021, p. 310) claims.[24]

4 The role of intuitions, remaining challenges, and methodological pluralism

So far, I have argued that theory-heavy proponents can adopt a credence-based theory-heavy approach that avoids the DOD and the SP. This is important because if the DOD and the SP are successful, they would undermine a theory's predictions about non-human consciousness even if we succeed in establishing a well-confirmed and complete theory of human consciousness. The purpose of this last section is as follows. First, I clarify the role that intuitions should play in the credence-based theory-heavy approach. Second, I will discuss some open challenges that the credence-based theory-heavy approach faces. Lastly, I will assess the possibility and dynamics of *methodological pluralism,* in which the credence-based theory-heavy approach forms a coalition with other methodological approaches.[25]

4.1 Clarifying the role of intuitions

It is always possible for theory-heavy proponents to simply dismiss intuitions, and embrace counterintuitive predictions (e.g., Murray, 2020; Tononi & Koch, 2015). Such an attitude might be motivated by the fact that there is a "broad range of antecedently plausible claims about the sparseness or abundance of consciousness in the world" (Schwitzgebel, 2020, p. 41), that it is plausible that the correct theory of consciousness defies common sense intuitions (Schwitzgebel, 2014), or that intuitions are simply untrustworthy (Murray, 2020). Theory-heavy proponents who dismiss intuitions can simply generate theory-based probability spaces and not worry about potential counterintuitive implications.

However, what about theory-heavy proponents who do want to take pre-theoretical intuitions about the distribution of consciousness seriously? Although intuitions are often appealed to in debates concerning the distribution question (e.g., Shevlin, 2021), what has been missing is a clarification of what intuitions *target.* Elucidating the target of intuitions also clarifies what role intuitions can play in the credence-based theory-heavy approach. A mistake that is often made is that intuitions are directed towards the wrong target. Recall that Shevlin (2021) uses counter-intuitive implications to argue against the conservative and liberal theory-heavy approaches. However, Shevlin argues that counterintuitive predictions indicate that there is something wrong with the methodologies (i.e., conservatism and liberalism) *themselves*. But this is a

[24] This is not to say that no efforts should be made to find the right level of specificity. But the credence-based theory-heavy approach shows that finding this holy grail is not necessary to make inferences about consciousness in non-human systems.

[25] The concept of methodological pluralism is distinct, and not to be confused, with the pluralistic stance.

mistake. To see why, suppose, for the sake of argument, that theories *do* necessitate and entail, for instance, a conservative approach. Now suppose that, by using the conservative approach, a theory makes wildly counter-intuitive predictions. What our pre-theoretical intuitions indicate, then, is that there is something wrong with the theory, not with the methodology that is used to extract implications from the theory. Hence, pre-theoretical intuitions target theories but not methodological approaches used to extract predictions from theories. Dismissing the use of a particular methodological approach requires more principled reasons. For instance, as I have argued for earlier, that they are simply not necessitated or entailed by the theory itself. This insight clarifies what role intuitions can play in the credence-based theory-heavy approach. Pre-theoretical intuitions *can* be used to dismiss or constrain a theory if there is a big discrepancy between intuition-based ascriptions and the outcomes in the theory-based probability-space that is generated by weighted theory-informed markers.[26]

4.2 Remaining challenges

4.2.1 The consensus problem

So far, on the credence-based theory-heavy approach, the likelihood of a target system being conscious was strictly determined by examining the probabilities associated with theory-informed markers. That is, we have antecedently accepted a particular theory to be true, and subsequently made a probability assessment about the presence of consciousness in a target system based on which kind of theory-informed marker it satisfies. However, should we not also factor in our antecedent credence in the theory itself? If our antecedent credence in a theory (e.g., GWT) is very low, finding evidence that a target system satisfies a theory-informed marker (e.g., a full-fledged global broadcast mechanism) should not move our credence that this system is conscious much. There is even reason to think that our credence in *all* available theories right now should be very low. There is no consensus as to which theory, *if any of them,* is right and the field notoriously faces a host of methodological obstacles (see e.g., Lau, 2022; Michel, 2019b). In the worst-case scenario, then, we have a very poor if not non-existent theoretical grasp on consciousness.[27] What does this imply for the credence-based theory-heavy approach?

If the worst-case scenario is true, we will simply have to conclude that it is too premature to use any theory-heavy approach to make inferences about non-human consciousness. After all, the theory-heavy methodology is only as good as the theories that are injected into it. However, the main contribution of this paper is to answer a more principled challenge that the theory-heavy approach faces. If Birch's DOD and Shevlin's SP are correct, we would still be incapable of making predictions about non-human consciousness *even if we have high antecedent credence in a theory.* When

[26] See Sect. 4.3 for a further discussion on the dynamics between other approaches to non-human consciousness and the credence-based theory-heavy approach.

[27] It is, of course, also possible to deny this and argue that we *should* have high credence in a particular theory. A proponent of a particular theory of consciousness will likely take this approach. I do not take a stance on the plausibility of any theory of consciousness. What I am concerned with is to map out possibilities and implications for the credence-based theory-heavy approach.

the time comes that such high antecedent credence is warranted, we better have a methodology ready to make predictions about non-human consciousness based on this theory. The credence-based theory-heavy approach offers precisely that.

However, perhaps one is more optimistic and considers it plausible that at least some theories are on the right track but is not yet willing to commit to any particular theory. In that case, one way to overcome the problem of low antecedent credence in a particular theory is by considering the predictions of multiple theories simultaneously.[28] For instance, one can start by describing a set of *front-runner* theories, for instance including, but not limited to, GWT (Baars, 1988; Mashour et al., 2020), integrated information theory (Tononi & Koch, 2015), higher-order theory (Rosenthal, 2005) and attention schema theory (Graziano, 2017).[29] Afterwards, one can analyze if this cluster of front-runner theories converge toward a consensus, or points towards a similar direction, with their predictions. In this case, even if our antecedent credence in each individual theory is fairly low, our credence with respect to the presence of consciousness in a target system can still be changed if a set of theories converge towards the same predictions about said system.[30] Whether such a consensus emerges is an open question, and in case it does not it "would be a disappointing result. But it is, inevitably, impossible to know in advance what the empirical outcome of a methodological strategy will be" (Birch, 2022, p. 145).

Moreover, it already seems useful to assess what theories would predict concerning the distribution question for other reasons. Drawing out a theory's implications for non-human consciousness need not only be used to make inferences but can also be used to reflect upon the theory itself. For instance, I argued earlier that pre-theoretical intuitions should target theories. Hence, those taking the *intuition-based* approach seriously should be interested in the predictions of a theory to determine whether there is a discrepancy between its predictions and pre-theoretical intuitions. If this is the case, this can be used against the theory itself (e.g., Aaronson, 2014).

Lastly, all previous considerations seem to presuppose a *winner-takes-all* view in which multiple theories compete to explain the same phenomenon. However, an alternative approach is to take a more unifying attitude towards theories of consciousness in which different theories simply capture different aspects, or dimensions, of consciousness (Ludwig, 2022). On this view, theories of consciousness are not in competition but instead deliver different perspectives on a same larger, perhaps poorly understood, phenomenon (see Wiese, 2018). Similarly, some recent theories of consciousness explicitly attempt to do justice to the multi-dimensional nature of consciousness (e.g., Newen & Montemayor, 2023). If this way of thinking is correct, different theories might produce predictions based on theory-informed markers about the presence and complexity of different dimensions of consciousness, instead of each producing predictions about the presence of consciousness as such. For instance, satisfying

[28] Chalmers (2023) has recently suggested something similar in the form of a so-called *theory-balanced* approach in which probabilistic predictions from multiple theories are considered.

[29] Chalmers (2023) suggests that surveying researchers in the field might be a way to determine what counts as a front-runner theory. It would be interesting to see if more objective criteria could be developed to determine what counts as a front-runner theory. See Seth and Bayne (2022) for a more in-depth review of the theoretical landscape.

[30] In a similar way, different theories may end up supporting the same kind of indicators (see Dung, 2022).

a full-fledged theory-informed marker makes it very likely that the dimension the theory tries to capture is part of the target system's conscious experience. Moreover, the complexity of capacities (e.g., attention) taken to be enhanced by the mechanism (e.g., an attention schema) described by a theory [e.g., attention schema theory (Graziano, 2017)] can give an indication how rich this dimension is for a target system. The credence-based theory-heavy approach, then, accommodates this unifying approach by generating a theory-based consciousness profile in which different features of consciousness, described by different theories, are scored in virtue of the presence of relevant theory-informed markers.[31] This approach, however, requires a radical revision in what theorists take their theory to be targeting; something which some theorists might not be willing to accept. But it is important to recognize that this option is available for theory-heavy proponents.

4.2.2 Calibrating the likelihoods of theory-informed markers

Another challenge is to find appropriate ways to adjust the likelihoods that are associated with theory-informed markers. The most straightforward way to do this is to use empirical insights. Setting aside some roadblocks to the measurements of consciousness (Browning & Veit, 2020; Michel, 2019b; Seth et al., 2008), empirical evidence can help calibrate likelihoods by, for instance, finding that a system is conscious despite lacking a *full-fledged* implementation of the candidate mechanism. For instance, suppose that someone with a severe brain injury, thereby lacking some features described in the full-fledged candidate mechanism, nevertheless *verbally reports* all the things you would expect of a conscious agent. In such a case, it becomes more likely that these additional features play no constitutive role for consciousness. Consequently, a more abstract interpretation of the theory that lacks these features, and accordingly a more coarse-grained theory-informed marker, can be assigned a higher likelihood.

It is also possible to use the so-called *no-report* paradigm to calibrate the likelihoods of theory-informed markers. No-report paradigms avoid the need for verbal reports by using experimental techniques such as "eye-movement, neuro-imaging, or physiological measures" (Duman et al., 2022). One particularly relevant advantage of this approach is that these techniques can be applied to systems that have a different architecture than humans (e.g., animals), yet cannot report on their conscious states (Tsuchiya et al., 2015). For instance, using a no-report paradigm, it *could* be shown that conscious contents can still be decoded despite the target system lacking a *full-fledged* implementation of the candidate mechanism. In such a case, again, a more abstract interpretation of the theory, and accordingly a more coarse-grained theory-informed marker, can be assigned a higher likelihood. Whether or not no-report paradigms can be successful is an ongoing debate (e.g., Block, 2019; Michel & Morales, 2020; Overgaard & Fazekas, 2016). However, the point here is not to take a stance in this debate. Rather, it is to show a potential avenue that theorists can exploit to calibrate the likelihoods associated with theory-informed markers. In the following section, I

[31] Such a proposal will be very much like Dung and Newen's (2023) consciousness profile. However, the consciousness profile of a credence-based theory-heavy approach explicitly bases its dimensions and markers on theories of consciousness.

will suggest another way in which likelihoods can be calibrated, namely by working together with alternative approaches to non-human consciousness.

4.3 Methodological pluralism

While the credence-based theory-heavy approach on its own can already be fruitfully applied, perhaps we should be careful in adopting a singular methodological attitude, where a preferred methodological approach is deemed to *compete* with other approaches. This runs the risk of undermining important insights that alternative approaches might offer. Thus, it is worth exploring whether different methodological approaches can be effectively synergized by embracing a form of *methodological pluralism*. Recently, Shevlin (2021) has provided a helpful framework, a so-called modest theoretical approach, that allows different approaches to benefit from each other. In what follows I will assess this modest theoretical approach in more depth. First, I will describe the view. Afterwards, I will (i) clarify its dynamics for the credence-based theory-heavy approach, (ii) suggest an extension for the modest theoretical proposal, and (iii) provide a motivation for the view.

Shevlin articulates the modest theoretical approach as follows:

> **Modest theoretical approach:** "The theory-heavy and theory-light approach[32] can operate in a form of dynamic equilibrium with insights of each informing and constraining the other (Shevlin, 2021, p. 308).

The dynamics of epistemic markers, as specified by the theory-light approach (Birch, 2022), and theory constraining, or informing, each other work as follows. Suppose we encounter a system that exhibits trace conditioning, reversal learning, and cross-modal learning behavior but lacks the cognitive mechanism dictated by our preferred theory of consciousness. In that case, we may have reason to "disregard the presence of the markers as reliable indicators" (Shevlin, 2021, p. 310). This would be an example of theory constraining epistemic markers. Similarly, suppose that "having used the markers to identify stronger and weaker consciousness candidates, we might assess whether a given theory can be coherently spelled out in such a way as to include the stronger candidates while excluding weaker ones" (Shevlin, 2021, p. 309). This would be an example of epistemic markers constraining or informing theory.

4.4 The dynamic equilibrium and the credence-based theory-heavy approach

How epistemic markers restrict and inform theory will be slightly different for the credence-based theory-heavy approach as it has thus far been presented. Shevlin argues that the role of epistemic markers is to clarify the *right* level of specificity of a theory (Shevlin, 2021, p. 310). This role for epistemic markers flows naturally from him taking the *singular* stance as I described earlier. However, by taking the

[32] Recall that a theory-light approach makes only minimal theoretical commitments, for instance that consciousness facilitates a cluster of cognitive abilities, including trace conditioning, reversal learning, and cross-modal learning (Birch, 2022, p. 140). These cognitive abilities subsequently act as behavioral epistemic markers to infer consciousness in non-human systems.

pluralistic stance, the role of epistemic markers is simply to help calibrate the likelihoods associated with theory-informed markers at various levels of specificity. For instance, suppose that a system strongly exhibits a cluster of epistemic markers such as trace conditioning, reversal learning, and cross-modal learning. Suppose also that these cognitive capacities can be selectively switched on and off under masking conditions the same way as in humans (Birch, 2022). However, suppose that this system implements a rather abstract version of the cognitive mechanism. In this case, there is reason to increase the likelihood that a system implementing a more abstract version of the cognitive mechanism is indeed conscious, since these cognitive capacities which are facilitated by consciousness in humans appear to be facilitated in the same way in our target system in virtue of only an abstract implementation of the proposed mechanism. Importantly, theorists will likely disagree about which cognitive capacities are facilitated by consciousness (Birch, 2022), because different theories will imply different associative links between their proposed mechanism and cognitive capacities. I take it, then, to be an important task for theorists to explicate which epistemic markers are compatible with their theory and use them in the aforementioned way. As such, the modest theoretical approach is another way to calibrate the likelihoods associated with theory-informed markers.

4.5 Generalizing Shevlin's modest theoretical approach

However, the modest theoretical approach, as described by Shevlin, seems too restrictive in the sense that it only describes dynamics that might occur between the theory-heavy and theory-light approaches. Yet it is also possible to establish these dynamics with other available approaches. For instance, one might not be convinced by the theory-light approach and instead might feel closer aligned with the theory-neutral approach. In that case, a dynamic equilibrium between the theory-heavy approach and the theory-neutral approach can be established. Moreover, as I have argued earlier, someone who takes the intuition-based approach seriously can also establish a dynamic equilibrium between the intuition-based and theory-heavy approach. Thus, a more inclusive description of the modest theoretical approach is appropriate:

> **Generalized modest theoretical approach:** The theory-heavy approach can couple with the theory-light, theory-neutral, or intuition-based approach to operate in a form of dynamic equilibrium with insights of each informing and constraining the other.

Which methodological approach a theory-heavy proponent decides to team up with depends, of course, on how convinced they are by arguments against the theory-neutral approach (e.g., Birch, 2022), theory-light approach (e.g., Halina et al., 2022; Schwitzgebel, 2020 and Shevlin, 2021), or intuition-based approach (e.g., Murray, 2020). In any case, plenty of options are available and this kind of methodological pluralism can fruitfully be explored.

4.6 Motivating the (generalized) modest theoretical approach

One apparent issue is that whilst Shevlin (2021, p. 310) discusses the benefits of the modest theoretical proposal, he does not motivate or justify why adopting this approach is reasonable in the first place. However, it is possible to legitimize Shevlin's proposal by drawing on Chang's (2004) analysis of measurements and scientific progress. Chang observes that there is a circularity between theories and observations because "empirical science requires observations based on theories, but empiricist philosophy demands that those theories should be justified by observations" (Chang, 2004, p. 221). To escape this circularity, Chang (2004) proposes that neither observations nor theories are self-justifying.[33] What happens instead is that observations and theories keep informing each other through iterative processes in which successive stages build upon, and correct, previous stages (Chang, 2004, p. 44; p. 226). That is, "we throw very imperfect ingredients together and manufacture something just a bit less imperfect" (Chang, 2004, p. 226). Chang's framework mirrors the dynamics of the modest theoretical proposal; neither epistemic markers nor theories of consciousness are self-justifying but rather inform each other to produce something *a bit less imperfect* over time through an iterative process. As such, adopting a modest theoretical approach seems reasonable since it mirrors the dynamics of scientific progress in other domains.

5 Conclusion

I have proposed a credence-based theory-heavy approach that incorporates a pluralistic stance and a probabilistic stance; theorists should generate multiple interpretations of their theory, develop associated *weighted* theory-informed markers, and use these markers to infer the presence of consciousness in target systems. The outcomes can subsequently be depicted in a theory-based probability space. The credence-based theory-heavy approach is capable of dealing with Birch's DOD and Shevlin's problem of specificity, both of which implicitly assume an inapt characterization of the theory-heavy approach. Hopefully, this positions the theory-heavy approach as a more compelling approach than it is currently being perceived as. Diffusing these challenges also allows us to turn our attention the other challenges the theory-heavy approaches currently faces, namely how to deal with the uncertain status of current theories of consciousness and how to calibrate weights associated with theory-informed markers. The credence-based theory-heavy approach is a tentative, but promising, proposal that theorists can explore within, or beyond, the space of possibilities I laid out in this paper.

Acknowledgements I would like to express my sincere gratitude to Albert Newen, Wanja Wiese, Leonard Dung and Fred Keijzer for their support, guidance, and insightful feedback. The quality of this paper has benefited greatly from it. I would also like to thank two anonymous referees for their valuable feedback. In addition, I would like to thank members of the reading group situated cognition from Ruhr-Universität Bochum (RUB) for helpful comments.

[33] Chang's views contrast with *foundationalism;* the view that certain beliefs can be self-justifying and thereby constituting one's evidence base (Foley, 1998, pp. 158–159).

Author contributions The manuscript "A Credence-based Theory-heavy Approach to Non-human Consciousness" was entirely written by myself.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author certifies that he has no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. Competing interests: The author certifies that there were no competing interests in the development of this manuscript.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

See Figs. 1, 2, 3.

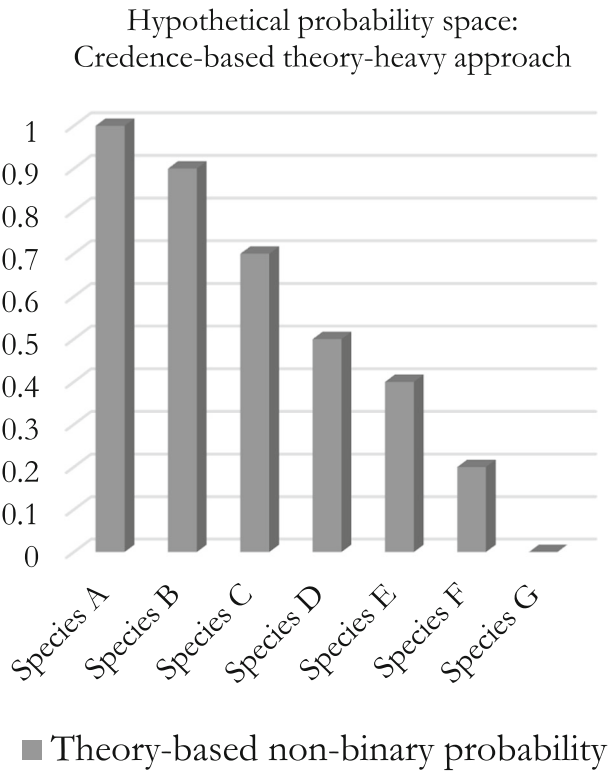

Fig. 1 A hypothetical theory-based non-binary probability space generated by a credence-based theory-heavy approach. Probabilities are based on the presence of theory-informed markers These theory-informed markers are operationalizations of different interpretations of a theory. Typically, if a system exhibits a more full-fledged theory-informed marker, it is more likely to be conscious

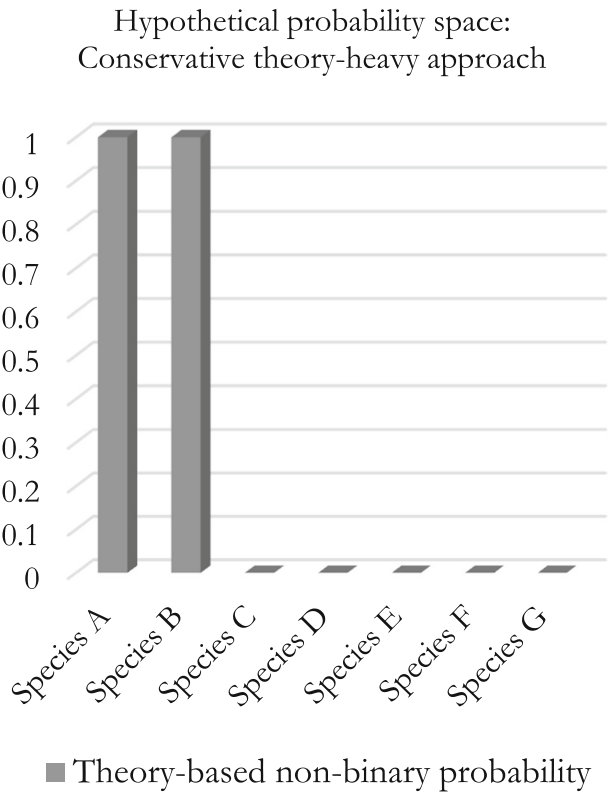

Fig. 2 A hypothetical theory-based binary probability space according to the conservatism response. Because the full range of capacities associated with the candidate mechanism are required, not many systems are predicted to be conscious. Moreover, the predictions are binary; a system is either deemed conscious or not

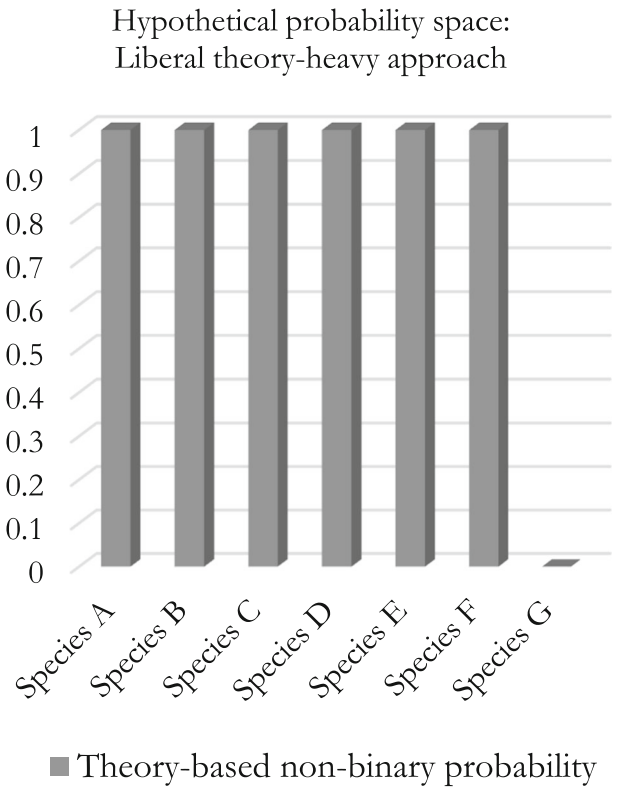

Fig. 3 A hypothetical theory-based binary probability space according to the liberalism response. Because the candidate mechanism is spelled out at the greatest possible degree of generality, many systems are predicted to be conscious. Moreover, the predictions are binary; a system is either deemed conscious or not

References

Aaronson, S. (2014, May 21). Why I am not an integrated information theorist (or, the unconscious expander). *Shtetl-Optimized*. https://www.scottaaronson.blog/?p=1799

Allen, C. (2016, October 24). Animal consciousness. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/consciousness-animal/

Allen, C., & Bekoff, M. (1997). *Species of mind: The philosophy and biology of cognitive ethology*. MIT.

Allen, C., & Bekoff, M. (2007). Animal consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 58–71). Blackwell. https://doi.org/10.1002/9780470751466

Andrews, K. (2024). "All animals are conscious": Shifting the null hypothesis in consciousness science. *Mind & Language*. https://doi.org/10.1111/mila.12498

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Birch, J., Schnell, A.K., & Clayton, N.S. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences, 24*(10), 789-801. https://doi.org/10.1016/j.tics.2020.07.007

Birch, J. (2022). The search for invertebrate consciousness. *Noûs, 56*(1), 133–153. https://doi.org/10.1111/nous.12351

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences, 18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences, 23*(12), 1003–1013. https://doi.org/10.1016/j.tics.2019.10.001

Browning, H., & Birch, J. (2022). Animal sentience. *Philosophy*. *Compass, 17*(5), e12822. https://doi.org/10.1111/phc3.12822
Browning, H., & Veit, W. (2020). The measurement problem of consciousness. *Philosophical Topics, 48*(1), 85–108. https://doi.org/10.5840/philtopics20204815
Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness*. http://arxiv.org/abs/2308.08708
Carruthers, P. (1999). Sympathy and subjectivity. *Australasian Journal of Philosophy, 77*(4), 465–482. https://doi.org/10.1080/00048409912349231
Carruthers, P. (2018a). Comparative psychology without consciousness. *Consciousness and Cognition, 63*, 47–60. https://doi.org/10.1016/j.concog.2018.06.012
Carruthers, P. (2018b). The problem of animal consciousness. *Proceedings and Addresses of the American Philosophical Association, 92*, 179–205.
Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*. Retrieved from https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking Press.
Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
Dennett, D. C. (1995). Animal consciousness: What matters and why. *Social Research, 62*, 691–710.
Duman, I., Ehmann, I. S., Gonsalves, A. R., Gültekin, Z., Van den Berckt, J., & van Leeuwen, C. (2022). the no-report paradigm: A revolution in consciousness research? *Frontiers in Human Neuroscience, 16*, 861517. https://doi.org/10.3389/fnhum.2022.861517
Dung, L. (2022). Assessing tests of animal consciousness. *Consciousness and Cognition, 105*, 103410. https://doi.org/10.1016/j.concog.2022.103410
Dung, L., & Newen, A. (2023). Profiles of animal consciousness: A species-sensitive, two-tier account to quality and distribution. *Cognition, 235*, 105409. https://doi.org/10.1016/j.cognition.2023.105409
Elwood, R. W. (2016). A single strand of argument with unfounded conclusion. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1056
Foley, R. (1998). Justification, epistemic. In E. Craig (Ed.), *Routledge encyclopedia of philosophy* (Vol. 5, pp. 157–165). Routledge.
Gennaro, R. I. (2004). Higher-order thoughts, animal. In J. Rocco & R. I. Gennaro (Ed.), *Higher-order theories of consciousness: An anthology*. John Benjamins, London.
Godfrey-Smith, P. (2016). Pain in parallel. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1057
Graziano, M. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI., 4*(60), 1–9. https://doi.org/10.3389/frobt.2017.00060
Halina, M., Harrison, D., & Klein, C. (2022). Evolutionary transition markers and the origins of consciousness. *Journal of Consciousness Studies, 29*(3), 62–77. https://doi.org/10.53765/20512201.29.3.077
Jamieson, D. (1998). Science, knowledge, and animal minds. *Proceedings of the Aristotelian Society, 98*, 79–102.
Key, B. (2016). Why fish do not feel pain. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1011
Klein, C., Hohwy, J., & Bayne, T. (2020). Explanation in the science of consciousness: From the neural correlates of consciousness (NCCs) to the difference makers of consciousness (DMCs). *Philosophy and the Mind Sciences*. https://doi.org/10.33735/phimisci.2020.II.60
Lau, H. (2022). *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press.
Ludwig, D. (2022). The functional contributions of consciousness. *Consciousness and Cognition, 104*, 103383. https://doi.org/10.1016/j.concog.2022.103383
Manzotti, R. (2016). No evidence that pain is painful neural process. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1034

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron, 105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Merker, B. H. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*, 63–81. https://doi.org/10.1017/S0140525X07000891

Michel, M. (2019a). Fish and microchips: On fish pain and multiple realization. *Philosophical Studies, 176*(9), 2411–2428. https://doi.org/10.1007/s11098-018-1133-4

Michel, M. (2019). Consciousness science underdetermined: A short history of endless debates. *Ergo*. https://doi.org/10.3998/ergo.12405314.0006.028

Michel, M., & Morales, J. (2020). Minority reports: Consciousness and the prefrontal cortex. *Mind and Language, 35*(4), 493–513. https://doi.org/10.1111/mila.12264

Murray, S. (2020). A case for conservatism about animal consciousness. *Journal of Consciousness Studies, 27*(9–10), 163–185.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435–450. https://doi.org/10.2307/2183914

Newen, A., & Montemayor, C. (2023). The ALARM theory of consciousness: A two-level theory of phenomenal consciousness. *Journal of Consciousness Studies, 30*(3), 84–105. https://doi.org/10.53765/20512201.30.3.084

Overgaard, M., & Fazekas, P. (2016). Can no-report paradigms extract true correlates of consciousness? *Trends in Cognitive Sciences, 20*(4), 241–242. https://doi.org/10.1016/j.tics.2016.01.004

Prinz, J. (2005). A neurofunctional theory of consciousness. In A. Brook & K. Akins (Eds.), *Cognition and the brain: The philosophy and neuroscience movement* (pp. 381–396). Cambridge University Press.

Prinz, J. (2018). Attention, working memory, and animal consciousness. In K. Andrews & J. Beck (Eds.), *The Routledge handbook of philosophy of animal minds* (pp. 185–195). Routledge/Taylor & Francis Group.

Rosenthal, D. (2005). *Consciousness and mind*. Oxford University Press.

Schwitzgebel, E. (2014). The Craziest Metaphysics of Mind. *Australasian Journal of Philosophy, 92*(4), 665-682. https://doi.org/10.1080/00048402.2014.910675

Schwitzgebel, E. (2020). Is there something it's like to be a garden snail. *Philosophical Topics, 48*(1), 39–63. https://doi.org/10.5840/philtopics20204813

Seacord, B. (2011). Animals, phenomenal consciousness, and higher-order theories of mind. *Philo, 14*(2), 201–222. https://doi.org/10.5840/philo201114214

Seth, A. K. (2016). Why fish pain cannot and should not be ruled out. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1038

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews. Neuroscience, 23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences, 12*(8), 314–321. https://doi.org/10.1016/j.tics.2008.04.008

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language, 36*(2), 297–314. https://doi.org/10.1111/mila.12338

Striedter, G. (2016). Lack of neocortex does not imply fish cannot feel pain. *Animal Sentience*. https://doi.org/10.51291/2377-7478.1037

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical transactions of the Royal Society of London Series B, Biological Sciences, 370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences, 19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Tye, M. (2016a). *Tense bees and shell-shocked crabs: Are animals conscious?* Oxford University Press.

Tye, M. (2016). Are insects sentient? *Animal Sentience*. https://doi.org/10.51291/2377-7478.1134

Veit, W. (2022a). Towards a comparative study of animal consciousness. *Biological Theory, 17*(4), 292–303. https://doi.org/10.1007/s13752-022-00409-x

Veit, W. (2022b). The origins of consciousness or the war of the five dimensions. *Biological Theory, 17*, 276–291. https://doi.org/10.1007/s13752-022-00408-y

Wiese, W. (2018). Towards a mature science of consciousness. *Frontiers in Psychology, 9*, 693. https://doi.org/10.3389/fpsyg.2018.00693

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness, 2020*(1), niaa013. https://doi.org/10.1093/nc/niaa013

Wiese, W. (2023). *Minimal models of consciousness: Understanding consciousness in human and non-human systems*. Unpublished manuscript. Retrieved from https://philpapers.org/rec/WIEMMO-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.