


Review

Tests for consciousness in humans and beyond

Tim Bayne ^{1,2,*} Anil K. Seth,^{2,3} Marcello Massimini,^{2,4,17} Joshua Shepherd,^{2,5,16} Axel Cleeremans,^{2,6} Stephen M. Fleming,^{2,7,14} Rafael Malach,^{2,8} Jason B. Mattingley,^{2,9} David K. Menon,^{2,10} Adrian M. Owen,^{2,11} Megan A.K. Peters,^{2,12} Adeel Razi,^{2,13,14} and Liad Mudrik^{2,15}

Which systems/organisms are conscious? New tests for consciousness ('C-tests') are urgently needed. There is persisting uncertainty about when consciousness arises in human development, when it is lost due to neurological disorders and brain injury, and how it is distributed in nonhuman species. This need is amplified by recent and rapid developments in artificial intelligence (AI), neural organoids, and xenobot technology. Although a number of C-tests have been proposed in recent years, most are of limited use, and currently we have no C-tests for many of the populations for which they are most critical. Here, we identify challenges facing any attempt to develop C-tests, propose a multidimensional classification of such tests, and identify strategies that might be used to validate them.

How is consciousness distributed?

There is a general consensus that healthy, awake, adult humans are conscious. Beyond that consensus, however, lies significant disagreement about the distribution of consciousness. There is debate about when consciousness first emerges in human development [1–6]; when it is retained (or regained) in the context of disorders of consciousness [7–9], such as the **unresponsive wakefulness syndrome (UWS)** (see [Glossary](#)) or the **minimally conscious state**, or in epileptic seizures [10–12]; and the degree to which it is present in sleep [13] and during anesthetic sedation [14,15]. There is also debate about the presence or possibility of consciousness in nonhuman animals [16–21], **neural organoids** [22–24], AI systems [25–29], and **xenobots** [30,31]. Is consciousness a relatively rare phenomenon that emerges late in ontogenesis, is rarely retained in conditions of non-responsiveness and severe brain damage, is restricted to only a few species, and cannot occur in organoids, AI systems, and/or xenobots? Or is consciousness more widespread, appearing early in development, retained even in some of the most severe forms of brain damage, found across numerous species, and capable of taking synthetic form?

To make progress here, scientists have proposed a number of tests for consciousness ('C-tests', [Figure 1](#); note that by using this term, we are not assuming a universal, decisive test, but instead refer to a battery of potential tests). A short and non-exhaustive list of proposed C-tests includes: the **command-following test** [32], the **narrative comprehension test** [33], the **sniff test** [34], the **perturbational complexity index (PCI) test** [35,36], the **P300/P3b global effect test** [37], the **AI consciousness test (ACT)** [27], and the **unlimited associative learning (UAL) test** [38,39]. Importantly, none of these tests has been offered as applicable to all contexts in which C-tests are needed and most have been offered only as possible tests of human consciousness. However, extending these tests to nonhuman animals, organoids, or artificial systems ([Figure 2](#); for extensions that have already been made, see [16,23]) raises questions about validation and the degree to which the test in question can be trusted.

Highlights

Developing validated tests for consciousness (C-tests) applicable to many different systems is a key challenge for consciousness science. We suggest a framework for doing so, highlighting fundamental challenges with respect to validation.

We propose a four-dimensional space within which potential C-tests can be positioned.

We suggest that a promising strategy is to focus on non-trivial human cases (infants, fetuses, disorders of consciousness) and then progress toward nonhuman systems (animals, artificial intelligence, neural organoids).

C-tests can inform and shape our understanding of consciousness and our evaluation of theories of consciousness.

¹Department of Philosophy, Monash University, Melbourne, VIC, Australia

²Canadian Institute for Advanced Research (CIFAR), Brain, Mind, and Consciousness Program, Toronto, ON, Canada

³Sussex Centre for Consciousness Science and School of Engineering and Informatics, University of Sussex, Brighton, UK

⁴Department of Biomedical and Clinical Science, University of Milan, Milan, Italy

⁵Universitat Autònoma de Barcelona, Bellaterra, Spain

⁶Center for Research in Cognition and Neuroscience, ULB Institute of Neuroscience, Université libre de Bruxelles, Brussels, Belgium

⁷Department of Experimental Psychology, University College London, London, UK

⁸The Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel

⁹Queensland Brain Institute and School of Psychology, The University of Queensland, Brisbane, QLD, Australia



Thus, there is a pressing need to develop a framework for C-tests, both to inform the application of existing C-tests and to help guide the development of new C-tests. This review proposes such a framework and employs it to illustrate some of the most pressing challenges in this space. We begin by considering what exactly a C-test might be and what might motivate the search for C-tests. We then turn to what it might mean to validate a C-test and how the process of validation might be conducted. While we advocate a specific approach to validation, our primary aim is to identify some of the critical decision points raised by the C-test project.

What is a C-test?

The central goal of any C-test is to determine whether a target system has subjective and qualitative experience – often called ‘phenomenal consciousness’. In other words, a good C-test should address the question of whether it ‘feels like’ something to be the target system [40]. While a subset of C-tests might focus on whether the target system has the capacity for phenomenal consciousness (which may not be currently realized), here we focus on tests that ask if a system is currently conscious.

The scientific importance of developing C-tests is clear, for knowing whether a particular kind of entity falls within the distribution of consciousness may have an important bearing on accounts of the nature of consciousness, such as whether it is necessarily unified [41], comes in degrees [42], or is multiply realized [43,44]. This rationale is particularly powerful if human experience is limited to a small, and perhaps idiosyncratic, region in the space of possible states of consciousness, as may be the case. (Arguably, trying to develop a comprehensive account of consciousness by studying only humans would be akin to trying to develop a comprehensive account of the elements by studying only copper.) The development of C-tests is also motivated by substantial societal and moral concerns, for there is broad consensus that consciousness has important implications for an entity’s **moral status** and in particular for how it ought to be treated (Box 1).

Box 1. The moral significance of consciousness

Across philosophy, bioethics, and medical ethics, many argue that the possession of phenomenal consciousness justifies the attribution of moral status to an entity (e.g., [80]). The interplay between attributions of consciousness and questions of moral status has been explored with reference to many domains, including nonhuman animals [81], cerebral organoids [82], AI [83,84], fetuses [85], psychiatric disorders [86], and traumatic brain injury [87]. While many agree that consciousness is morally significant, however, there is little agreement as to why this is the case [88].

A popular view, often called sentientism (or ‘narrow sentientism’) [89], is that a capacity for valenced experiences – those that feel good or bad – is required for moral status. A minority view (sometimes called ‘broad sentientism’) holds that the capacity to have any experiences at all is sufficient for moral status [90]. A third view, not in direct competition with the first two, holds that the moral significance of consciousness comes in degrees, and that the overall complexity or sophistication of an entity’s experiences mediates the attached level of moral status [91,92]. A fourth view maintains that moral status is closely connected to self-consciousness rather than phenomenal consciousness *per se* [93].

Debates about the moral significance of consciousness bear on the development of C-tests in several ways. Most obviously, different views regarding the moral significance of consciousness will make different C-tests seem morally urgent, and that in turn will impact issues such as media coverage, funding priorities, and decisions about what kinds of tests to emphasize (e.g., those that focus on the relative complexity of conscious states as opposed to those that focus merely on the capacity for some kind of pleasure or pain). Further, the moral urgency of a particular C-test may intersect with the stringency of any proposed validation of that C-test and with the threshold that ought to be set for what counts as a statistically significant result [44]. Whether we are content with some level of validation for a test may depend on the moral significance of the target population, for example, or the moral significance of the aspects of consciousness the C-test targets. Given the high moral stakes, some stakeholders (e.g., animal rights advocates, AI developers, policy advocates, laypeople) may demand only low degrees of confidence in a measure and/or a prioritization of sensitivity over specificity, an approach that is often defended with reference to the ‘precautionary principle’ [88,94,95].

¹⁰University of Cambridge, Cambridge, UK

¹¹University of Western Ontario, London, ON, Canada

¹²University of California, Irvine, Irvine, CA, USA

¹³Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia

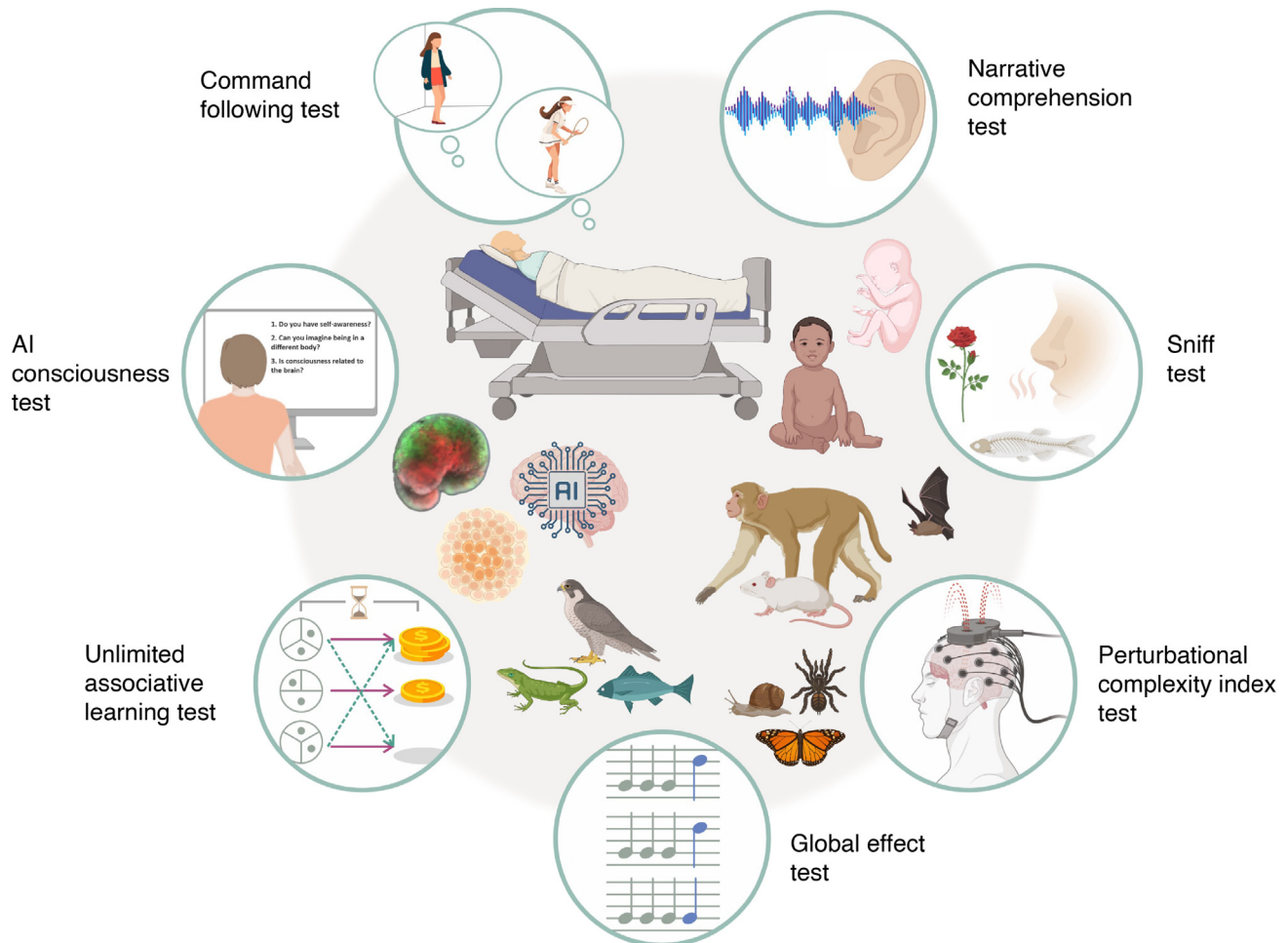
¹⁴Wellcome Centre for Human Neuroimaging, University College London, London, UK

¹⁵School of Psychological Sciences and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

¹⁶ICREA, Barcelona, Spain

¹⁷IRCCS Fondazione Don Gnocchi

*Correspondence:
tim.bayne@gmail.com (T. Bayne).



Trends in Cognitive Sciences

Figure 1. Examples of proposed tests for consciousness (C-tests). The surrounding circles illustrate several of the C-tests in use today, while the center of the figure features some of the populations for which C-tests are needed: non-responsive patients, babies, fetuses, nonhuman animals, artificial intelligence (AI) systems, neural organoids, and xenobots. From top left, clockwise. Command-following test [32]: While their brain activity is recorded, the individual is instructed to imagine either playing tennis or walking through their home. The test is passed if the individual's brain activity is comparable to that of healthy controls. Narrative comprehension test [33]: While their brain activity is recorded, the target is presented with a short segment of a movie (or its soundtrack). The test is passed if the neural activity indicates executive processing akin to that seen in healthy controls. 'Sniff' test [34]: The target is presented with pleasant/aversive odorants and their sniffing response is used to classify levels of consciousness. Perturbational complexity index (PCI) test [35,36]: A transcranial magnetic stimulation (TMS) pulse is applied to the target and electroencephalography (EEG) is used to measure the complexity of the neural response to that pulse. The test is passed if the target's level of complexity is comparable to that found in behaviorally responsive individuals. P300/P3b 'global effect' test [37]: Several sequences of sounds are presented one after another; in the first sequences, the last sound in the sequence always diverges from the previous one, creating local (or first-order) divergence. Critically, in the last sequence, no local divergence appears, creating global (or second-order) divergence only. Detection of this global divergence is taken as evidence for consciousness. Unlimited associative learning (UAL) test [38,39]: The test is passed if the target system learns the relations between compound, novel stimuli (here, circles with dots at different locations), separately in time, so that learning involves trace conditioning (represented here by the clock), whose values can be adapted and reversed given different circumstances (represented here by the broken arrows). Note that UAL also involves second-order conditioning, which for simplicity is not depicted here. The AI consciousness test (ACT) [27]: A (boxed) AI system is taken to have passed the ACT test if it appears to spontaneously understand and use concepts about internal experiences.

Targeting phenomenal experience means that C-tests ought not be primarily directed at the detection of intelligence [45], self-regulation, voluntary behavior, or any other capacity that might be associated with consciousness [46]. Of course, if it can be shown that a certain capacity (e.g., attention [47], perceptual organization [48]) covaries with consciousness – as some **theories of consciousness** suggest – then it might well be fruitful for a C-test to focus on

		Command following	Narrative	Sniffing	PCI	Global effect	ACT	UAL
ALTERED STATES	Sedation	+	+	+	+	+	?	+
	Epileptic seizure	+	+	+	+	+	?	+
	Sleep/Dreaming	+ ?	+ ?	+ ?	+	+ ?	?	+ ?
UNCLEAR CAPACITY FOR CONSCIOUSNESS	Disorders of consciousness	+	+	+	+	+	?	+
	Babies	—	?	+	+	+	—	+
	Fetuses	—	—	?	+	+	—	?
NON-HUMAN ANIMALS	Non-human mammals	—	—	+	+ ?	+	—	+
	Non-mammal vertebrates	—	—	—	—	+	—	+
	Invertebrates	—	—	—	—	?	—	+
ARTIFICIAL SYSTEMS	Neural organoids	—	—	—	?	+ ?	—	+ ?
	xenobots	—	—	—	?	+ ?	—	+ ?
	AI	+ ?	+ ?	—	?	+ ?	+	+ ?

Trends in Cognitive Sciences

Figure 2. The scope of tests for consciousness (C-tests). The applicability of different C-tests (rows) to different populations (columns), divided into levels based on the provisional order of validation suggested here. Plus (+) signs indicate that a test can likely be administered to a specific population in a meaningful way (although its specificity/sensitivity might be low), possibly with some modifications. Dashes (—) denote the inapplicability or irrelevance of the test for a specific population. Question marks (?) denote that the test might be applicable but more development is needed to test whether this is the case. Finally, a combination of a plus sign and a question mark (+?) signifies that although the test can be applied it is unclear what its results would mean. See Box 3 for an explanation of what it means for populations to be stratified into levels.

that capacity as a means of identifying phenomenal consciousness. Even then, however, the primary goal of the C-test would be the detection of consciousness as such, and the detection of the related capacity would merely be a means to that end.

Different types of C-tests approach the detection of consciousness in different ways. One class of C-tests focuses on the generic property of being conscious and would provide little to no information about the system's conscious contents. For example, some C-tests target features that putatively reflect general properties of conscious experience, such as neural integration and differentiation ([35]; for a review, see [49]). Another class of C-tests focuses on specific contents of consciousness or on psychological capacities held to be sufficient for consciousness. These include bodily sensations (e.g., pain, smell [50]), voluntary responsiveness (e.g., producing imagery following a command to do so [32]), a dissociable response to a subliminal versus a supraliminally presented stimulus [51], and learning of various kinds [17,37]. The interpretation of these tests as C-tests depends, of course, on the assumption that the tested capacity/content is indeed a strong indicator of consciousness.

C-tests: a multidimensional space

There are many dimensions along which C-tests differ, and these dimensions could be mapped in different ways. Here, we develop a four-dimensional C-test space that we consider illuminating (Figure 2), particularly with respect to questions of validation (see the following section).

Dimension 1: target population

One dimension along which C-tests can differ concerns the kind(s) of populations that the test targets. The strongest possible test would have an unrestricted domain and would be applicable

to any kind of system. However, such a universal test might be unattainable, and many C-tests, particularly current and near-future ones, will be applicable only to particular populations. For example, some C-tests might apply only to humans, only to mammals, only to evolved biological systems, or only to AI. The optimal way of applying C-tests and demarcating relevant populations remains unclear; we return to this issue in the last section of this review.

Dimension 2: specificity

A second dimension concerns the specificity of a C-test: how low its false-positive rate is. An example of a C-test with high specificity (in humans) is the command-following test [32], which is assumed to be a reliable indicator of consciousness in many clinical contexts – an embedded assumption in the diagnostic criteria for the minimally conscious state [52]. Although command following is sometimes observed in the context of epileptic absence seizures [11], it is unclear whether such patients are fully unconscious [10,53]. Moreover, the command following seen in epileptic seizures appears to be restricted to overlearned, motor behavior (e.g., walking, clenching a fist), which is arguably different from the kind of command-following used in C-tests.

Importantly, a test's specificity is population dependent. For example, command following might be highly specific in humans but not in (certain kinds of) AI systems. Thus, we cannot assume that C-tests that are highly specific in one population will be highly specific in another. We suggest that even those procedures that have relatively low levels of specificity for a given population might be legitimately regarded as 'C-tests', for the information that they provide might still be meaningful, 'shifting the dial' on questions relating to the distribution of consciousness.

Dimension 3: sensitivity

Sensitivity refers to how well the test identifies true positives – that is, systems that are genuinely conscious. A C-test with high sensitivity will have a low proportion of false negatives. Notably,

Box 2. Types of validity

Cognitive science distinguishes various types of validity [96,97], each of which is relevant to the search for C-tests. Construct validity concerns whether the test measures the phenomenon of interest; in this case, phenomenal consciousness. An important variant of construct validity is discriminant validity: how exclusive is the test? That is, does it probe consciousness itself or distinct, albeit related, phenomena? Determining whether a test has construct validity is particularly challenging for consciousness, for it will often be difficult to tell whether a test locks on to consciousness itself or to some other phenomenon (e.g., reportability).

Second, content validity – does the test probe for every possible aspect of the target phenomenon? Here, a test that is high on content validity will tell us not only whether the target is conscious but what kinds of conscious states/contents it is in (e.g., whether it is in pain). Whether content validity is required depends on the aim of the C-test. Content validity is not required if the goal is to simply detect the presence/absence of consciousness, although if certain contents are essential to consciousness then information about those aspects will be relevant to the detection of consciousness. If, however, our aim is to determine not just whether there is something it is like to be the target system, but also what it is like to be that system, then content validity is required.

Third, criterion validity – what is the predictive accuracy of the test with respect to various outcomes? Here, a crucial question concerns which outcomes are relevant to criterion validity. It is fairly clear what ought to count as a 'relevant outcome' in certain populations. For example, when C-tests are applied to patients with disorders of consciousness, it would be reasonable to treat a patient's recovery into full consciousness as a relevant outcome [35,50]. However, there are many other populations (e.g., infants, nonhuman animals, AI systems) for which the specification of a 'relevant outcome' is less straightforward.

Finally, face validity – do the results of the test converge with intuitive, folk-psychological judgments about the distribution of consciousness, or does the test generate verdicts that are radically at odds with those judgments? Here, there are important questions both about what face validity involves ('whose intuitions are we talking about?') [58,59,98] and the degree to which it should be respected (see further discussion with respect to the iterative NK strategy; Box 4).

Glossary

AI consciousness test (ACT): a test for consciousness in (boxed) AI systems based on the system's capacity to spontaneously develop and appropriately use consciousness-related concepts.

Command-following test: the (covert) command-following test probes for consciousness in behaviorally non-responsive patients based on their capacity to produce neural activity indicative of command-appropriate mental imagery.

Deflationary conceptions of consciousness

consciousness: deflationists hold that the nature of consciousness can be fully understood *a priori*, whereas non-deflationists hold that understanding the nature of consciousness requires empirical investigation.

Global workspace theory (GWT): a theory of consciousness that posits that consciousness occurs when information enters a global workspace which in humans (and potentially other animals) is mediated by specialized neurons, primarily located in the frontoparietal cortex.

Integrated information theory (IIT): a theory of consciousness that identifies consciousness with the unfolded cause-effect structure specified by the system in a certain state.

Minimally conscious state: a disorder of consciousness in which patients show minimal, unstable, yet clear signs of being aware of some aspects of their environment. These signs may manifest as intentional behavior, language comprehension, or emotional responses.

Moral status: an entity has moral status if and only if its interests have intrinsic (or non-derivative) moral significance.

Multiple realizability: the claim that the same mental state can be instantiated in systems with different physical/computational/architectural properties.

Narrative comprehension test: this test looks for the presence of consciousness in behaviorally non-responsive patients by probing their capacity to exhibit neural activity indicative of narrative comprehension.

Natural kind: a term describing the grouping of entities that reflects the structure of the natural world (and thus supports inference and generalization) rather than mere human interest.

specificity and sensitivity need not be correlated. For example, the sensitivity (in patients) of the command-following test [32] is presumably lower than its specificity, for there are many reasons why a patient might fail to follow a command despite being conscious (e.g., they fail to hear or comprehend the command). This is the mirror image of the situation discussed above (concerning AI), in which the test is arguably highly sensitive but not highly specific.

Both sensitivity and specificity are of crucial importance. C-tests that are highly specific but with low/unknown sensitivity do not allow us to make warranted statements about the absence of consciousness. Equally, highly sensitive C-tests with low/unknown specificity fall short in determining the presence of consciousness. Either way, using one of these two classes of tests in isolation would fail to capture the true spectrum of consciousness.

Dimension 4: rational confidence

This dimension positions C-tests according to the degree to which our estimates of their sensitivity and specificity are warranted. At one end of this spectrum are tests that have the kind of rational confidence associated with ordinary, folk-psychological judgments about the presence/absence of consciousness in (neurotypical, adult) humans. This is obviously desirable, but it seems unlikely to be obtainable in the near future, especially with respect to C-tests that are applied to populations that are very different from us. At the other end of the spectrum are tests that evoke lower rational confidence and are best treated as merely suggestive. In between, of course, lie a range of intermediate positions. For example, we might view estimates of a test's specificity/sensitivity as reasonable given current evidence, but such estimates may change as further data are acquired.

Determining how 'rational confidence' ought to be assigned requires an account of how putative C-tests might be validated. We now turn to that topic.

Strategies for validating a C-test

The toughest challenge for any attempt to develop robust C-tests concerns validation: how do we identify the specificity and sensitivity of a test? After all, we have no independent way of assessing whether the members of the target population are conscious – and if we did, we would not need C-tests.

Below, we consider three possible validation strategies. These strategies are not exclusive (a single C-test could be defended on the basis of more than one strategy), but we engage with them independently because each has a distinctive rationale, as we now explain.

The redeployment strategy

This involves arguing that the test under consideration is a variant of a C-test that already has widespread legitimacy (the 'established test'). Going back to the covert command-following test [32], one might argue that because overt (i.e., behavioral) command following is widely accepted as providing evidence of consciousness in patients, covert command following (i.e., command following that is not behaviorally evident) should also be treated as providing evidence of consciousness in this population [32,54,55].

Although the redeployment strategy has an important role to play in validating C-tests, it also has significant limitations. First, it cannot justify extending a test beyond the populations in which the established C-test operates. (As we previously noted, it would be implausible to treat command following as reliable evidence for the presence of consciousness in AI systems.) Second, it is inherently conservative; it has nothing to say in defense of our established consciousness-ascribing

Neural organoids: 3D cell culture models that are derived from animal stem cells and are aimed at mimicking some aspect of neural development, structure, or function.

P300/P3b global effect test: this test takes neural responses (which typically occur 300 ms after presentation of the relevant stimulus) to global (that is, second order) oddballs as a marker for consciousness. It has been applied to behaviorally non-responsive patients and infants/fetuses.

Perturbational complexity index (PCI) test: a test for consciousness that involves stimulating the brain with transcranial magnetic stimulation and measuring the complexity of the evoked electrocortical responses using electroencephalography.

Sniff test: developed in connection with behaviourally non-responsive patients, the sniff test involves presenting pleasant and aversive odors to a patient, and ascribing consciousness on the basis of a relative decrease in nasal inhalation volume relative to the presentation of plain air.

Theories of consciousness: theories that propose to identify the physical/functional basis of consciousness, explaining both what distinguishes conscious states or systems from unconscious ones and what distinguishes different kinds of conscious states from each other.

Unlimited associative learning (UAL) test: developed in connection with nonhuman animals, the UAL test for consciousness takes the capacity for unlimited associated learning (roughly, the flexible and reversible learning of compound, novel stimuli), to be an indicator of consciousness.

Unresponsive wakefulness syndrome (UWS): formerly/also known as the 'vegetative state'; UWS patients show intermittent periods of wakefulness but do not manifest behaviors suggestive of conscious awareness.

Xenobots: synthetic proto-organisms made from animal cells in novel configurations that exhibit spontaneous morphogenesis and behavior.

practices but takes their legitimacy for granted. Thus, the redeployment strategy does not provide a positive case for treating command following as evidence of consciousness. (That said, we would note that it is unclear whether any such defense is needed.) To provide such a case, one would need to either provide independent evidence linking command following to consciousness or appeal to a **deflationary conception of consciousness**, arguing that ‘by definition’ any system that can reliably follow commands is conscious.

Here, it is worth noting that any attempt to validate a C-test inevitably raises questions about folk conceptions of the distribution of consciousness (and a C-test’s face validity; [Box 2](#)). At one end of the spectrum, one might hold that no test that runs strongly counter to our ordinary attitudes (i.e., has poor face validity) could turn out to be warranted [\[56\]](#). At the other end, one might argue that folk-psychological attitudes ought to be afforded little weight in the validation process [\[57–59\]](#) and that a C-test could turn out to be strongly warranted even if it was at odds with many of our most deeply held assumptions about the distribution of consciousness. Between these two positions lie a number of more moderate attitudes that could be taken toward folk-psychological conceptions of the distribution of consciousness.

The theory-based strategy

A second strategy for validating a putative C-test appeals to the resources of a theory of consciousness. The idea here is that the ‘fit’ between a well-grounded theory of consciousness and a putative C-test provides us with a reason to treat that C-test as a reliable indicator of consciousness.

Appeals to theory-based considerations are not uncommon in discussions of C-tests. For example, the P300/P3b global effect test [\[37\]](#) has looked to the **global workspace theory (GWT)** for support [\[46\]](#), while the initial inspiration for the PCI test [\[34\]](#) came from the **integrated information theory (IIT)** of consciousness [\[60\]](#). Although any account of validation may need to appeal to theoretical considerations at some point, the theory-based validation strategy faces three significant challenges.

First, no single theory – or even theoretical framework – currently commands general assent [\[61\]](#). At least 22 distinct theories of consciousness are taken seriously [\[62\]](#) and some of these theories have variants that differ significantly from each other. Although many of these theories might lend their support to the same suite of C-tests [\[63\]](#), others clearly differ in terms of which C-tests they support; further, we can expect that these differences will be more pronounced the more ‘alien’ (i.e., unlike us) a population is. The existence of multiple theories of consciousness might not be problematic if the trend was toward integration and convergence, but that seems not to be the case; indeed, theories of consciousness appear to be proliferating [\[64\]](#). To address this problem, one might advance an integrative framework in which the results of multiple C-tests are combined, with the weighting of each test determined by the support given (by the consciousness science community) to the particular theory that lies behind it [\[65\]](#). Although this approach provides a useful mechanism for integrating the results of several C-tests, it may not yield broad agreement about the legitimacy of any particular C-test.

A second problem with the theory-based strategy is that many theories of consciousness are formulated with reference to (neurotypical, adult) humans, and it is often unclear how they ought to be applied to other populations (the ‘generalization problem’). For example, GWT makes detailed claims about the mechanisms of human consciousness, but it does not (yet) provide an account of what it would take for any kind of system to have a (consciousness-supporting) global

workspace. Although the core tenets of a theory may provide some guidance as to how it ought to be applied to alien populations [25,66], it seems likely that uncertainty regarding a theory's implications will increase the further we move beyond its primary evidential base (i.e., adult, neurotypical human beings).

Third, the theory-based strategy faces the threat of circularity: any theory of consciousness must itself be justified, but it seems as though the process of justifying a theory must itself appeal to C-tests. (After all, validating a theory of consciousness requires evaluating its predictions regarding the distribution of consciousness, and that process would seem to require C-tests.) If we appeal to the same theory of consciousness to validate those tests, we would seem to be reasoning in a circle, assuming that a theory can both justify and in turn be justified by a set of C-tests.

There are two ways in which one might respond to this challenge. First, one might argue that it is possible to validate a theory of consciousness without appealing to (contested) C-tests. Instead (the thought is), all we need to do is appeal to the predictions that the theory makes about 'consensus cases'. Although this proposal would indeed defuse the threat of circularity (if successful), we have doubts about whether it can be made to work. This is primarily because we would still need to justify the extrapolation of such predictions to non-consensus cases. Second, one might allow that validating a theory of consciousness does indeed require an appeal to (contested) C-tests, and thus accept that a certain kind of circularity is inherent in theory-based validation, but argue that this circularity might be virtuous rather than vicious. The idea here is that if the development of both theory and C-tests proceeds hand-in-hand, then a theory of consciousness can be used to validate a range of C-tests, and those C-tests can, in turn, play a foundational role in validating that theory. The next section explores an approach to validation that expands on this proposal in more detail.

The iterative natural kind (NK) strategy

A third strategy for validating a putative C-test involves treating consciousness as a natural kind [67,68]. A natural kind category is a category that cuts the world 'at its joints', grouping items based on commonalities in their underlying nature (as opposed to superficial similarities, arbitrary conjunctions of features, or mere human interests) [69–72]. If consciousness is a natural kind, then conscious systems will share an underlying nature, and in principle that nature should be identifiable via the kinds of iterative procedures that have succeeded in uncovering the underlying nature of other natural kinds, such as heat (see below). Here, we start with a large number of pre-theoretical measures of consciousness (judgments of face validity; Box 2) and revise and correct those judgments when prompted by consideration of theoretical unification, simplicity, explanatory power, and predictive success. This process guides the construction of (tentative, provisional) C-tests, which are in turn used to revise our conception of the distribution of consciousness. Circularity, here, becomes embedded in and defused by the iterative nature of theory development and testing. Although the validity of any putative C-test begins with pre-theoretical judgments, it is not simply derived from those judgments but can outstrip them in various ways. Pre-theoretical measures of consciousness (e.g., verbal report, command following) play a crucial role in the initial stages of our inquiry, but we regard them as open to revision following discoveries about the underlying natural kind. A useful parallel is provided by the history of thermometry: although the development of thermometers was catalyzed by our intuitive, folk-psychological judgments of temperature, those intuitive, folk-psychological judgments are now corrigible on the basis of readings taken by the thermometers that have been developed ([73]; see also [74,75]). Importantly, face validity (Box 2) is retained, for our pre-theoretical judgments regarding temperature (boiling water is hotter than ice) are generally in accord with the verdicts of our thermometers.

The iterative NK strategy also has the tools to address the generalization problem that troubles the theory-based approach. If we can identify an underlying feature (or features) that accounts for the pre-theoretical measures of consciousness with which we began, we would have a principled basis for generalizing beyond ‘consensus cases’ (roughly, human beings who are capable of producing reports) and applying our C-tests to the members of other populations (Box 3).

There are two further points to make about the iterative NK strategy. First, putative C-tests should be extended to novel populations by bootstrapping. The idea here is that C-tests must first be validated in ‘neighboring’ populations before being applied to more ‘alien’ populations, because the closer a population is to ‘us’, the more credence we should have in our pre-theoretical measures of consciousness. Of course, as we move away from ‘neighboring’ populations and toward populations that are increasingly ‘alien’, our faith in those pre-theoretical measures may begin to fade,

Box 3. A hierarchical approach to validation

An important feature of the iterative NK strategy is that it imposes a hierarchical structure on the validation process. Here, various C-tests are first validated in a ‘consensus population’ (level 1) in which we have agreed-upon ways of identifying the presence of consciousness (first and foremost, report). At this stage (Stage 1), all C-tests are well-correlated and are accordingly assigned relatively high credences. Those credences form the basis for priors in applying these C-tests to neighboring populations (level 2). At Stage 2, we test the correlations between these (previously validated) C-tests and other (new) C-tests that might be relevant to these populations [but not applicable at level 1; in the figure, test 5 (T5)]. In the scenario illustrated by Figure 1, population P2 exhibits high pairwise correlations between T1, T2, and T4 (but not between T1 and T3 or between any other test and T5); population P3 exhibits high pairwise correlations between all tests other than T1 and T5; and population P4 exhibits high pairwise correlations between all tests other than T3 and T5. In light of these results, Stage 3 involves updating credence in the C-tests, not only for level 2 populations but also for the level 1 population (‘consensus cases’). These results also inform our priors when applying these C-tests to the next level of the hierarchy (level 3).

A critical question concerns how populations should be assigned to levels. Here, different approaches can be taken. For example, one could measure proximity to the level 1 population based on cognitive and executive capacities, cognitive architecture, neurophysiology, or phylogenetic relations. Measuring proximity in one way would place human neonates closer to the consensus cases than adult octopuses, whereas measuring it in another way might reverse that order, placing adult octopuses closer to ‘us’ than our newborn conspecifics. Ideally, decisions about how to measure proximity between populations would not be made on *a priori* grounds but would be informed by empirical investigations into the nature of consciousness.

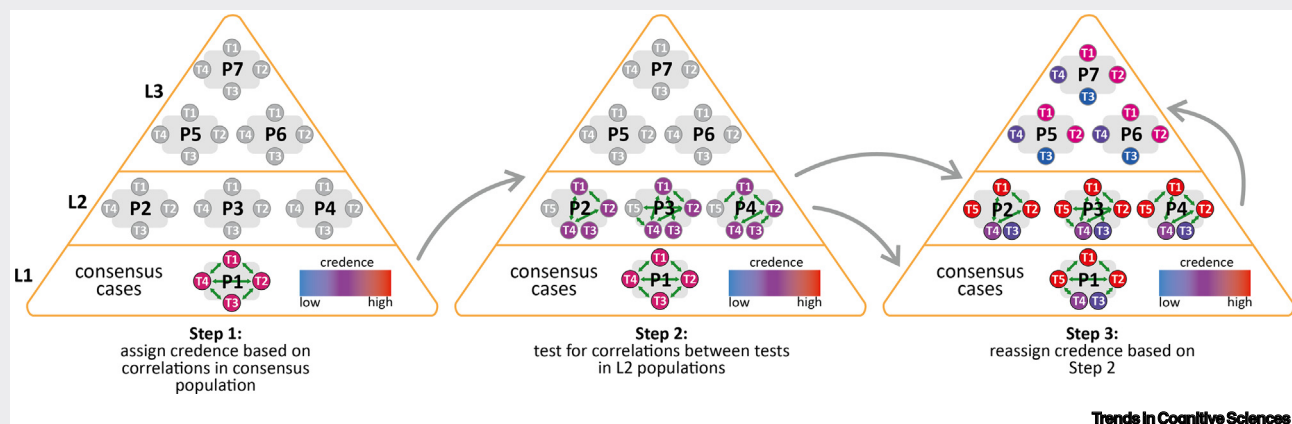


Figure 1. An illustration of the iterative natural kind (NK) strategy. Different populations (gray rectangles) are divided into levels (rows in the pyramids) hierarchically (within each pyramid, the further one is from the bottom, the consensus population, the higher the level). The figure portrays the iterative NK strategy, starting from Stage 1 (left) and progressing rightwards. At Stage 1, we assign relatively high credences to tests 1–4 (T1–T4) (marked in circles; colors depict the credence level; see the scale on the right) due to them being well correlated in the consensus population (correlations are depicted by arrows connecting the tests). At Stage 2, we rely on the credences from Stage 1 [level 1 (L1)] to serve as priors for level 2 (L2) populations. We then test the correlations between the tests (T1–T4 and a new test, T5, which was not applicable to the consensus population). Based on the results of these tests, in Stage 3 credences are reassigned with respect to both the L2 populations (bottom curved arrow) and the consensus population L1 (top curved arrow). The updated credences of these tests now become the starting point for the validation of tests applied to level 3 (L3) populations.

but we will now have increased confidence in those C-tests that have been validated in neighboring populations.

In embracing this hierarchical approach to validation, we leave open the important question of which populations are ‘closest’ to the consensus population, for proximity can be measured in various ways (e.g., behaviorally, functionally, neurophysiologically). Neurotypical individuals in global states characterized by behavioral non-responsiveness (e.g., sleep, sedation), humans with clinical conditions characterized by (some degree of) behavioral non-responsiveness (e.g., disorders of consciousness), human infants (and fetuses), and nonhuman animals that are close relatives of ours might each be treated as ‘neighboring populations’, for different reasons. At the same time, some populations (organoids, xenobots, AI systems) are clearly less close to the consensus cases than any of the above. We suspect that questions regarding how populations ought to be stratified will become clearer as we learn more about the mechanisms underpinning consciousness in ‘us’ (Box 3).

A second point is that C-tests cannot be validated in isolation from each other, but ought to be calibrated against one another [76]. After all, if two tests do indeed probe consciousness (i.e., if they have construct validity; Box 2), then their results ought to agree. The importance of employing independent tests derives from the fact that although any particular test will be subject to error, different tests are *a priori* less likely to be subject to the same kinds of errors.

Of course, challenging questions will arise when an individual fails one C-test while passing others. For example, a behaviorally non-responsive patient might pass one type of mental imagery test but not another [77], a nonhuman animal might show evidence of unlimited associative learning but fail to be responsive to second-order oddballs, or a human infant might be responsive to second-order oddballs but fail to pass tests of neural complexity. How should we respond in such cases?

The framework developed in the first half of this review provides useful first steps in addressing this question. First, we need to ascertain whether failure to pass a C-test can be attributed to that test not being applicable to the candidate system or being applied inappropriately. That is, is it the kind of test whose ‘failure’ is informative?

Second, we need to recognize that failing a particular test might tell us something about the target individual, but it might equally tell us something about the sensitivity/specificity of the relevant test. Consider the behaviorally non-responsive patient who engages in mental imagery on command (T1) but fails to show evidence of narrative comprehension (T2). Should we conclude that T1 is a false positive (and that the patient is not conscious) or that T2 is a false negative? Here, we can draw on Bayesian considerations [78,79], asking which conclusion best accords with our background assumptions regarding the sensitivity/specificity of T1 and T2 in this (and neighboring) populations.

In practice, of course, taking a Bayesian approach is challenging given the difficulties involved in coming up with well-specified priors, but it nonetheless has the advantage of forcing one to be transparent about prior beliefs and their justification, which may help to guide the application of the iterative NK strategy in general.

Concluding remarks

This review has sketched a multidimensional framework for addressing the challenges facing the quest to develop robust and general C-tests. By capturing the key features of a C-test (its target population, specificity, and sensitivity and the rational confidence associated with it), this

Outstanding questions

How can existing C-tests be extrapolated to nonhuman cases, where: (i) it is even harder to validate them in the absence of any ground truth; and (ii) some of the abilities required by the test are human specific (e.g., having a language, showing a specific pattern of neural activity)?

To what extent should C-tests converge with folk-psychological assumptions about the distribution of consciousness?

To what extent should C-tests be tied to existing theories of consciousness, and should we favor a theory-independent test, given lack of agreement about current theories?

Will it ever be possible to develop a single, universal C-test or will we be limited to a battery of different C-tests that apply to different aspects of consciousness and different kinds of systems?

framework not only enables us to grasp the strengths and weaknesses of individual C-tests, but also facilitates an understanding of how various C-tests are related to each other and how they may usefully be combined. Such a framework could serve as a basis for the development of new, better C-tests. Ideally, we would be able to develop a C-test that had universal applicability (i.e., could be applied to all populations), had perfect specificity and sensitivity, and was such that we were fully warranted in treating it as having universal applicability and perfect specificity/sensitivity (see [Outstanding questions](#)). However, even if it is possible to develop a ‘universal C-test’, that goal seems likely to require the development of a number of less comprehensive C-tests – that is, C-tests that apply to only a restricted range of populations, have less-than-maximal specificity and/or sensitivity, and whose status as C-tests is not beyond debate.

In addition to providing this framework, we have also confronted the challenging question of how C-tests might be validated. Here, we have examined three possible validation strategies: the redeployment strategy, the theory-based strategy, and the iterative NK strategy. In the absence of a theoretical consensus concerning the nature of consciousness, we believe that the latter represents the most promising of these approaches, although we also recognize that it faces significant challenges ([Box 4](#)). Crucially, we suggest that the field should adopt a bootstrapping approach in which the validation of C-tests begins with populations closest to ‘us’ (e.g., neurotypical individuals in global states characterized by behavioral non-responsiveness) and only then is extended to more challenging populations in an incremental fashion. This approach does not require that a C-test be fully validated in humans and nonhuman animals before it can be applied to an AI system or a neural organoid, but it does mean that our consensus population (e.g., neurotypical adult humans) retains a certain kind of priority when it comes to validation.

Over and above the specific claims we have defended here, our main aim has been to highlight the various decision points confronting the C-test project. These points include questions regarding the concept of consciousness, the reliability of folk-psychological assumptions about the

Box 4. Challenges to the NK strategy

Although the NK strategy is widely embraced in the sciences (in particular, the biological sciences), a number of theorists have argued that it cannot be successfully applied to consciousness [99]. Some have argued that because there is significant disagreement about pre-theoretical measures of consciousness, the approach is unlikely to lead to convergence on a single underlying kind [100] and, consequently, on a single set of C-tests. Other theorists have rejected the NK strategy on the grounds that the concept of consciousness is not a natural kind concept – that it lacks the structure required by the NK strategy [12]. A third objection is that empirical studies suggest that there is no unitary kind associated with consciousness [101].

However, perhaps the most pressing objection concerns the possibility that consciousness is multiply realized, so that the natural kind associated with consciousness in humans might be fundamentally different from the natural kind associated with consciousness in other kinds of systems [61,102]. The NK strategy does not accommodate that possibility very well, for, in starting with the markers of consciousness in ‘us’ (i.e., human beings who are uncontroversially conscious) the approach is set up to privilege tests of ‘human-like’ experience. Thus, it struggles to identify tests for radically alien forms of consciousness. The worry, in short, is that the NK strategy is unacceptably anthropocentric.

Although this challenge is certainly serious, a number of responses to it are possible. First, given the rejection of deflationary conceptions of consciousness, to reach any conclusions about the distribution of consciousness one must embrace some kind of hierarchical and iterative approach to validation that begins with consensus cases. Second, it might be argued that we do not (as yet) have empirical evidence that consciousness is in fact associated with multiple natural kinds, and that, until we do, worries about **multiple realization** can be safely set to one side [103,104]. Third, although there are open questions about how far beyond the consensus cases the iterative refinement of theories, measures, and tests might eventually take us, some optimism on this front can be drawn from the significant theoretical [49] and empirical [63] convergence on complexity-related measures of global states of consciousness in humans. The predictive success of these measures across many different conditions, such as sleep, anesthesia, hallucinatory states, coma, and related disorders, suggests that treating consciousness as a natural kind might already be yielding dividends [23].

distribution of consciousness, and the possibility of reaching consensus on a general theory of consciousness in the absence of a validated C-test and sufficient empirical data. These questions are clearly some of the most challenging for the science of consciousness; equally, they are also some of the most important.

Acknowledgments

The authors gratefully acknowledge support from the 'Brain, Mind, and Consciousness' program of the Canadian Institute for Advanced Research (CIFAR). We are also grateful for the feedback received from two anonymous referees for this journal, and the Editor.

Declaration of interests

M.M. is cofounder and shareholder of Intrinsic Powers.

References

- Moser, J.F. *et al.* (2021) Magnetoencephalographic signatures of conscious processing before birth. *Dev. Cogn. Neurosci.* 49, 100964
- Bayne, T. *et al.* (2023) Consciousness in the cradle: on the emergence of infant experience. *Trends Cogn. Sci.* 27, 1135–1149
- Kouider, S. *et al.* (2013) A neural marker of perceptual consciousness in infants. *Science* 340, 376–380
- Lagercrantz, H. and Changeux, J.P. (2010) Basic consciousness of the newborn. *Semin. Perinatol.* 34, 201–206
- Padilla, N. and Lagercrantz, H. (2020) Making of the mind. *Acta Paediatr.* 109, 883–892
- Passos-Ferreira, C. (2023) Are infants conscious? *Philos. Perspect.* 22, 308–329
- Naci, L. *et al.* (2017) Detecting and interpreting conscious experiences in behaviorally non-responsive patients. *Neuroimage* 145, 304–313
- Naccache, L. (2018) Minimally conscious state or cortically mediated state? *Brain* 141, 949–960
- Hermann, B.A. *et al.* (2021) Importance, limits and caveats of the use of "disorders of consciousness" to theorize consciousness. *Neurosci. Conscious.* 2021, niab048
- Blumenfeld, H. (2005) Consciousness and epilepsy: why are patients with absence seizures absent? *Prog. Brain Res.* 150, 271–286
- Gloor, P. (1986) Consciousness as a neurological concept in epileptology: a critical review. *Epilepsia* 27, S14–S26
- Ali, F. *et al.* (2012) The assessment of consciousness during partial seizures. *Epilepsy Behav.* 23, 98–102
- Windt, J.M. (2019) Can a microdynamic approach to sleep-onset imagery solve the overabundance problem of dreaming? Commentary on Tore Nielsen's "Microdream neurophenomenology". *Neurosci. Conscious.* 2019, niz005
- Nilsen, A.S. *et al.* (2022) Are we really unconscious in "unconscious" states? Common assumptions revisited. *Front. Hum. Neurosci.* 16, 987051
- Sanders, R.D. *et al.* (2012) Unresponsiveness ≠ unconsciousness. *J. Am. Soc. Anesthesiol.* 116, 946–959
- Birch, J. (2022) The search for invertebrate consciousness. *Nous* 56, 133–153
- Ginsburg, S. and Jablonka, E. (2019) *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*, MIT Press
- Carruthers, P. (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*, Oxford University Press
- Dung, L. and Newen, A. (2023) Profiles of animal consciousness: a species-sensitive, two-tier account to quality and distribution. *Cognition* 235, 105409
- Edelman, D.B. and Seth, A.K. (2009) Animal consciousness: a synthetic approach. *Trends Neurosci.* 32, 476–484
- Tye, M. (2016) *Tense Bees and Shell-shocked Crabs: Are Animals Conscious?*, Oxford University Press
- Lavazza, A. and Massimini, M. (2018) Cerebral organoids: ethical issues and consciousness assessment. *J. Med. Ethics* 44, 606–610
- Bayne, T. *et al.* (2020) Are there islands of awareness? *Trends Neurosci.* 43, 6–16
- Niikawa, T.Y. *et al.* (2022) Human brain organoids and consciousness. *Neuroethics* 15, 5
- Dehaene, S. *et al.* (2017) What is consciousness, and could machines have it? *Science* 358, 486–492
- Seth, A. (2021) *Being You: A New Science of Consciousness*, Penguin
- Schneider, S. (2020) How to catch an AI zombie: testing for consciousness in machines. In *Ethics of Artificial Intelligence* (Laio, M., ed.), pp. 439–458, Oxford University Press
- Butlin, P. *et al.* (2023) Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv*, Published online August 17, 2023. <https://doi.org/10.48550/arXiv:2308.08708>
- Dung, L. (2023) Tests of animal consciousness are tests of machine consciousness. *Erkenntnis*, Published online November 14, 2023. <https://doi.org/10.1007/s10670-023-00753-9>
- Blackiston, D.E. *et al.* (2021) A cellular platform for the development of synthetic living machines. *Sci. Robot.* 6, eabf1571
- Kriegman, S.D. *et al.* (2020) A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1853–1859
- Owen, A.M. *et al.* (2006) Detecting awareness in the vegetative state. *Science* 313, 1402
- Naci, L. *et al.* (2014) A common neural code for similar conscious experiences in different individuals. *Proc. Natl. Acad. Sci. U. S. A.* 111, 14277–14282
- European Parliament and Council of the European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *OJEU*
- Casartotto, S. *et al.* (2016) Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann. Neurol.* 80, 718–729
- Casali, A.G. *et al.* (2013) A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5, 198ra105
- Bekinschtein, T.A. *et al.* (2009) Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci. U. S. A.* 106, 1672–1677
- Birch, J. *et al.* (2020) Unlimited associative learning and the origins of consciousness: a primer and some predictions. *Biol. Philos.* 35, 56
- Ginsburg, S. and Jablonka, E. (2020) Consciousness as a mode of being. *J. Conscious. Stud.* 27, 148–162
- Nagel, T. (1974) What is it like to be a bat? *Philos. Rev.* 83, 435–450
- Bayne, T. (2010) *The Unity of Consciousness*, Oxford University Press
- Lee, A.Y. (2023) Degrees of consciousness. *Nous* 57, 553–575

43. Putnam, H. (1967) Psychological predicates. In *Art, Mind, and Religion* (Capitan, W.H. and Merrill, D.D., eds), pp. 37–48, University of Pittsburgh Press
44. Polger, T.W. (2006) *Natural Minds*, MIT Press
45. Turing, A.M. (1950) Computing machinery and intelligence. *Mind* 59, 433–460
46. Juliani, A. et al. (2022) On the link between conscious function and general intelligence in humans and machines. *arXiv*, Published online March 24, 2022. <https://doi.org/10.48550/arXiv:2204.05133>
47. Mashour, G.A. et al. (2020) Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798
48. Lamme, V.A.F. (2020) Visual functions generating conscious seeing. *Front. Psychol.* 11, 83
49. Sarasso, S. et al. (2021) Consciousness and complexity: a concision of evidence. *Neurosci. Conscious.* 7, 1–24
50. Arzi, A. et al. (2020) Olfactory sniffing signals consciousness in unresponsive patients with brain injuries. *Nature* 581, 428–433
51. Ben-Haim, M.S. et al. (2021) Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (*Macaca mulatta*). *Proc. Natl. Acad. Sci. U. S. A.* 118, e2017543118
52. Giacino, J.T. et al. (2002) The minimally conscious state: definition and diagnostic criteria. *Neurology* 58, 349–353
53. Bayne, T. (2011) The presence of consciousness in absence seizures. *Behav. Neurol.* 24, 47–53
54. Owen, A.M. et al. (2007) Response to comments on “Detecting awareness in the vegetative state”. *Science* 315, 1221
55. Monti, M.M. et al. (2010) Willful modulation of brain activity in disorders of consciousness. *N. Engl. J. Med.* 362, 579–589
56. Papineau, D. (2002) *Thinking About Consciousness*, Clarendon Press
57. Francken, J. et al. (2021) An academic survey on theoretical foundations, common assumptions and the current state of the field of consciousness science. *Neurosci. Conscious.* 2021, niac011
58. Arico, A. et al. (2011) The folk psychology of consciousness. *Mind Lang.* 26, 327–352
59. Jack, A.I. and Robbins, P. (2012) The phenomenal stance revisited. *Rev. Philos. Psychol.* 3, 383–403
60. Tononi, G. et al. (2016) Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461
61. Shevlin, H. (2021) Non-human consciousness and the specificity problem: a modest theoretical proposal. *Mind Lang.* 36, 297–314
62. Seth, A. and Bayne, T. (2022) Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452
63. Farisco, M. and Changeux, J.P. (2023) About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neurosci. Conscious.* 2023, niad016
64. Yaron, I. et al. (2022) The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* 6, 593–604
65. Chalmers, D.J. (2023) Could a large language model be conscious? *arXiv*, Published online March 4, 2023. <https://doi.org/10.48550/arXiv:2303.07103>
66. VanRullen, R. and Kanai, R. (2021) Deep learning and the global workspace theory. *Trends Neurosci.* 44, 692–704
67. Block, N. (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–499
68. Shea, N. and Bayne, T. (2010) The vegetative state and the science of consciousness. *Br. J. Philos. Sci.* 61, 459–484
69. Bird, A. and Tobin, E. (2012) Natural kinds. In *The Stanford encyclopedia of philosophy* (Zalta, E.N., ed.), Stanford University
70. Khalidi, M.A. (2023) *Natural Kinds*, Cambridge University Press
71. Hawley, K. and Bird, A. (2011) What are natural kinds? *Philos. Perspect.* 25, 205–221
72. Franklin-Hall, L.R. (2015) Natural kinds as categorical bottlenecks. *Philos. Stud.* 172, 925–948
73. Chang, H. (2004) *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press
74. Tal, E. (2013) Old and new problems in philosophy of measurement. *Philos. Compass* 8, 1159–1173
75. Tal, E. (2020) Measurement in science. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.), Stanford University
76. Michel, M. (2023) Calibration in consciousness science. *Erkenntnis* 88, 829–850
77. Gibson, R.M. et al. (2014) Multiple tasks and neuroimaging modalities increase the likelihood of detecting covert awareness in patients with disorders of consciousness. *Front. Hum. Neurosci.* 8, 950
78. Seth, A.K. and Dienes, Z. (2017) The value of Bayesian statistics for assessing credible evidence of animal sentience. *Anim. Sentience* 2, 22
79. Gelman, A. and Shalizi, C.R. (2013) Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66, 8–38
80. Kahane, G. and Savulescu, J. (2009) Brain damage and the moral significance of consciousness. *J. Med. Philos.* 34, 6–26
81. DeGrazia, D. (1996) *Taking Animals Seriously: Mental Life and Moral Status*, Cambridge University Press
82. Sawai, T. et al. (2022) Mapping the ethical issues of brain organoid research and application. *AJOB Neurosci.* 13, 81–94
83. Liao, S.M. (2020) The moral status and rights of artificial intelligence. In *Ethics of Artificial Intelligence* (Liao, M., ed.), pp. 480–503, Oxford University Press
84. Sinnott-Armstrong, W. and Conitzer, V. (2021) How much moral status could artificial intelligence ever achieve? In *Rethinking Moral Status* (Liao, M., ed.), pp. 269–289, Oxford University Press
85. Harman, E. (2003) The potentiality problem. *Philos. Stud.* 114, 173–198
86. Bayne, T. (2002) Moral status and the treatment of dissociative identity disorder. *J. Med. Philos.* 27, 87–105
87. Fernández-Espejo, D. and Owen, A.M. (2013) Detecting awareness after severe brain injury. *Nat. Rev. Neurosci.* 14, 801–809
88. Shepherd, J. (2023) Non-human moral status: problems with phenomenal consciousness. *AJOB Neurosci.* 14, 148–157
89. Lee, A.Y. (2019) Is consciousness intrinsically valuable? *Philos. Stud.* 176, 655–671
90. Chalmers, D.J. (2022) *Reality+: Virtual Worlds and the Problems of Philosophy*, Penguin
91. Shepherd, J. (2018) *Consciousness and Moral Status*, Taylor & Francis
92. Kreitmair, K. (2023) Consciousness and the ethics of human brain organoid research. *Camb. Q. Healthc. Ethics*, Published online March 23, 2023. <https://doi.org/10.1017/S0963180123000063>
93. Bermúdez, J.L. (1996) The moral significance of birth. *Ethics* 106, 378–403
94. Birch, J. (2017) Animal sentience and the precautionary principle. *Anim. Sentience* 2, 1
95. Birch, J. and Browning, H. (2021) Neural organoids and the precautionary principle. *Am. J. Bioeth.* 21, 56–58
96. Shepard, L.A. (1993) Evaluating test validity. *Rev. Res. Educ.* 19, 405–450
97. Wainer, H. and Braun, H.I. (2013) *Test Validity*, Routledge
98. Robbins, P. (2006) The phenomenal stance. *Philos. Stud.* 127, 59–85
99. Bayne, T. and Shea, N. (2020) Consciousness, concepts and natural kinds. *Philos. Top.* 48, 65–83
100. Phillips, I. (2018) The methodological puzzle of phenomenal consciousness. *Philos. Trans. R. Soc. B Biol. Sci.* 373, 20170347
101. Irvine, E. (2012) *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*, Springer
102. Block, N. (2002) The harder problem of consciousness. *J. Philos.* 99, 391–425
103. Cao, R. (2022) Multiple realizability and the spirit of functionalism. *Synthese* 200, 506
104. Polger, T.W. (2009) Evaluating the evidence for multiple realization. *Synthese* 167, 457–472