Leonard Dung - cite the published version: https://doi.org/10.1016/j.concog.2022.103410

Assessing tests of animal consciousness

1. Introduction

In this entire paper, the term 'consciousness' is used as a shorthand for 'phenomenal consciousness', i.e., the subjective experiential feeling of being in certain mental states. The underlying question is: Which animal species have consciousness? To find out which animals have conscious experiences, we have to make inferences from their behavior and from the cognitive capacities this behavior manifests.[1] This raises the question: Which kinds of behavior provide strong evidence that an animal is conscious? Moreover, we may want to know which behaviors provide at least weak evidence in contrast to behaviors which provide negligible or no support for attributions of consciousness. There seems to be no agreement on this fundamental issue, even among researchers who are sympathetic to the notion that consciousness is widespread in the animal kingdom. For a vast set of different tests of animal consciousness has been proposed and different researchers emphasize competing indicators.

I will briefly illustrate this diversity. According to one strain of thought, consciousness can best be measured by testing for certain learning capacities (Birch 2022; Ginsburg and Jablonka 2019). By contrast, the literature on animal pain emphasizes indicators like preferences for analgesics, long-term behavioral changes in response to bodily damage and motivational trade-offs between competing demands (Sneddon et al. 2014). Tye (2017), in his extensive treatment of consciousness in fish and insects, counts features like the susceptibility to perceptual illusions, the capacity for transitive inference or the presence of negative judgement bias as evidence of consciousness. Further putative cognitive-behavioral indicators of consciousness include attention and working memory (Prinz 2018), goal-directed action selection (Butlin 2020), metacognition (Shea and Heyes 2010), posttraumatic stress disorder (Hoffman 2020) and feats such as tool-use, kin recognition and long-term memory (Brown 2015).

In this paper, I aim to formulate guidelines for assessing such tests of consciousness. This paper motivates and systematizes multiple desiderata for tests of animal consciousness which have been proposed in the literature. Moreover, these desiderata can also be used to assess sets of tests of animal consciousness and to shed light on when and how different tests

[1] The presence of certain neural structures or processes may also constitute evidence of animal consciousness. Currently, due to challenges in fruitfully exploiting neural data (Crump and Birch 2022), progress on animal consciousness is mainly driven by investigating cognitive and behavioral properties. However, I will mention neuroscientific knowledge as additional constraint, where it is relevant.

complement each other. This list of desiderata helps to explain what it is that makes certain tests, and sets of tests, of animal consciousness compelling. Thereby, one can rely on this list to assess whether the indicators mentioned or any future tests of animal consciousness which will be proposed provide sufficient evidence of consciousness. Moreover, one can design future tests of animal consciousness with the intention of satisfying these desiderata to provide confidence of their evidential strength.

In developing this set of desiderata, this paper makes three novel contributions: First, the literature contains many scattered discussions of putative tests of consciousness which are mostly considered in isolation. This paper brings important proposals from the literature in relation to each other such that a general and coherent "checklist" for putative tests of animal consciousness can be formed. Second, while the literature often divides proposed tests according to dichotomous categories like 'valid' and 'invalid', distinguishing different desiderata allows more nuanced judgements which emphasize differing strengths and weaknesses of particular consciousness tests. Third, this paper provides reasons for accepting certain desiderata from an encompassing bird's eye point of view. While isolated proposals for tests of animal consciousness and reasoning in favor of them abound, they are typically not viewed from a general perspective which evaluates them according to a consistent list of considerations.

I will divide the desiderata for tests of animal consciousness into three categories. I will begin next section by suggesting desiderata which capitalize on putative analogies between indicators of consciousness in humans and other animals. For each desideratum, I will explain why it increases the strength of a test of animal consciousness and what its limits are. In addition, I will offer an example of current animal consciousness research in which this desideratum is respected. In section 3 and 4, I will proceed accordingly with desiderata which are supported by the main theories of the nature of consciousness and desiderata which can further enhance tests which fulfill desiderata belonging to the other two categories. In section 5, I will apply the list of desiderata to evaluate tests of animal consciousness from the literature and to suggest refinements.

2. Arguments from analogy

It is generally agreed that there is no definitive evidence for or against attributions of consciousness, at least once we set aside adult human beings and only consider beings which lack the ability for flexible, untrained and language-based reports of experiences. Different pieces of evidence can be more or less strong, such that – for different animals – we should be

more or less confident that consciousness is present. This encourages us to think about the distribution of consciousness in a nuanced manner, meaning that we assign different credences to beliefs of consciousness in various species and update these credences in accord with every new piece of evidence. If we stick to this framing, then this paper discusses which tests of consciousness should, if they are passed, strongly increase our degree of belief in consciousness of the species in question.

Now, I begin examining which desiderata tests of consciousness have to satisfy to warrant significant updates of those credences. Crucially, concrete empirical tests which fulfill these desiderata provide positive evidence that an animal is conscious. However, animals may fail to pass single specific empirical tests inspired by those criteria for all kinds of reasons which do not imply that they are not conscious. Most notably, all desiderata I propose are either based on the human-likeness of an animal's behavioral responses or aim to reveal its high-level cognitive capacities. Animals may fail many tests based on these desiderata because they lack the necessary intelligence or because their cognitive-behavioral repertoire is unlike what we are familiar with from humans. Nevertheless, this does not speak against attributions of consciousness as strongly as passing those tests would favor attributions of consciousness. Thus, the focus is on positive, not negative, markers of animal consciousness (Ginsburg and Jablonka 2019).

The first category of desiderata for consciousness tests ultimately rests on arguments from analogy to human consciousness and its behavioral expression. By satisfying these desiderata, tests of consciousness gain legitimacy because passing them involves exhibiting behavior or cognitive processes which are causally enabled by consciousness in humans. Thus, these desiderata are justified in virtue of *Newton's Principle* (NP) which was introduced into the literature by Michael Tye. The idea is the following: "The general point is that I am entitled to infer sameness of cause from sameness of effect […], *unless I have evidence that defeats the inference*" (Tye 2017, p. 74).

That is, given that one event has a particular cause, one is entitled to infer that an event of the same type possesses a cause of the same type, unless there is overriding evidence to the contrary. Applied to the investigation of animal consciousness, NP entails that one is justified to prefer the hypothesis that an animal is conscious over its negation if the animal performs behavior which is caused by conscious experience in humans and there is no overriding evidence suggesting that the animal is not conscious. One influential objection to the use of NP concerns the question of what counts as defeating or overriding evidence. It has been argued that we need extensive prior knowledge about the distribution of consciousness to specify

defeaters and that we need to be able to rule out defeaters to apply NP (Birch 2022; Michel 2019). This would compromise the utility of NP. An adequate defense of NP is beyond the scope of this paper. Nevertheless, I note that the demand to rule out potential defeaters before applying NP is too strong. An inference to the best explanation can be valid, despite being defeasible. The same goes for inferences via NP. Inferences based on NP can support attributions of animal consciousness, even if there is a chance that a defeater is present.

Inferences which are based on analogies between human and animal behavior are typically justified by reference to shared biological similarity and evolutionary history between humans and the animals in question. The idea is that the overall biological and evolutionary similarity supports the claim that similar behavior is caused by similar processes. A corollary is that such inferences are weaker regarding animals which are evolutionarily distant from humans, such as invertebrates. As I see it, the evidential strength of analogies between human and other animal behavior is a function of both their overall biological similarity and the closeness of the specific analogy. Thus, in order to persuade us via evidence stemming from the use of NP that invertebrates are conscious, more striking evidence is needed than for ensuring us that mammals are conscious. But this does not mean that analogy evidence is useless in the invertebrate case. We have to look for particularly striking behavioral and cognitive similarities.[2]

	Nr.	Desideratum	References
Analogy	1a	Behavior which is shared with humans and not a simple reflex.	(Godfrey-Smith 2016a; Sneddon et al. 2014)
	1b	Behavior which, according to credible human studies, depends on consciousness in humans.	(Droege et al. 2021; Tye 2017)
	1c	Double dissociation which mirrors the double dissociations found between human conscious and unconscious perception.	(Ben-Haim et al. 2021; Crump and Birch 2021)
Theoretical integration	2a	Evidence of selective attention	(Gabay et al. 2013; Krauzlis et al. 2018; Prinz 2018)

[2] I have the same reply to the "measurement problem of consciousness" characterized by Browning and Veit (2020). This problem is also based on the insight that types of behavior which are caused by conscious experience in humans do not need to be caused by conscious experience in all organisms. While the problem is real, strengthening the analogy between human and non-human behavior – while also finding evidence which is supported by theories of consciousness (see later desiderata) – gradually strengthens the case for attributions of non-human consciousness.

	2b	Evidence of widespread and complex information integration	(Butlin 2020; Irvine 2020)
	2c	Evidence of metacognition	(Perry and Barron 2013; Yuki and Okanoya 2017)
Paradigms from human consciousness science	3a	The stimuli used are masked (or subject to similar experimental manipulations validated in human studies).	(Birch 2022)
	3b	There are systematic correlations between putative indicators of consciousness.	(Birch 2022; Shea and Bayne 2010)

Table 1. This table summarizes the eight desiderata for tests of animal consciousness championed in this paper. The left column indicates that they divide into three main categories. The right column lists references to theoretical discussion and empirical application of the respective desideratum.

Desideratum 1a states that responses to stimuli provide better evidence that the stimulus was perceived consciously if they are human-like and (relatively) non-reflexive. A behavior is "human-like" if it superficially matches the behavior humans perform in the same type of situation. For instance, if an animal responds to a noxious stimulus like a human who was hurt (e.g., avoidance of the stimulus and subsequent limping) and this response is not a reflex, then this suggests that the animal may be in conscious pain. This is based on the insight that there is a class of "simple reflexes" which do not indicate consciousness. Withdrawal reflexes in response to nociception, e.g. pulling your hand away from a hot stove, are an example. This is demonstrated by the fact that these behaviors are initiated unconsciously in humans. Evidence for this is that noxious stimulation sometimes causes humans in vegetative state to cry out, withdraw or display pain-suggesting facial expressions (Laureys 2007). In addition, the lower limbs of complete spinal cord patients, in which the patients cannot feel anything, still exhibit the withdrawal flexion reflex (Dimitrijević and Nathan 1968; Key 2016, p. 4). Moreover, withdrawal reflexes are quick and automatic and are typically mediated by neurons in the spinal cord and brainstem. There is no reason to expect such behaviors to require consciousness. To the extent that behavioral responses transcend simple reflexes, they are better evidence of consciousness.[3]

[3] Of course, desideratum 1a does not entail that every behavior which is not a simple reflex is accompanied by consciousness, as (e.g.) unconscious modulation of nociception shows (Mason and Lavery 2022).

The paradigm of looking for non-reflexive human-like responses to noxious stimuli is largely what guides current research on animal pain. Behaviors are non-reflexive to the extent that they are flexible, instead of stereotypic, and involve cognitive processes like memory, planning and representation of the external world. In this research area, a diverse set of indicators is used to test for animal pain (for overviews, see Sneddon (2014) and Crump et al. (2022)). They include preferences for analgesics, self-administration of analgesia, trade-off behavior, place preference or avoidance conditioning and long-term behavioral changes after bodily damage. The unifying feature of this diverse list is that all its elements refer to non-reflexive responses to noxious stimuli. Since these responses are rather complex and involve behavior which we see as indication of conscious pain when it occurs in humans, they are a first step in developing reliable tests of animal consciousness.

However, the limit of this approach is that proponents normally do not provide systematic evidence for the assumption that the non-reflexive behavior in question is not caused by unconscious processes. This is a problem because unconscious processes are sufficient even for many non-reflexive, relatively flexible behaviors like driving (in sleepwalking) or precise grasping movements (when guided by the dorsal visual stream) (Carruthers 2020). For this reason, mere intuition cannot reveal which human behaviors depend on consciousness. We need to scrutinize these assumptions empirically. Only this way we can have justified confidence that a type of behavior is not possible without consciousness in humans which would make its presence in an animal good evidence of consciousness.

Based on these reflections, desideratum 1b demands that the proposed test of animal consciousness should be validated on humans. That is, it should be provided evidence that the feature which is taken as an indicator of animal consciousness depends on consciousness in humans. Trace conditioning – a form of classical conditioning in which the presentation of the unconditioned and the conditioned stimulus is separated by a temporal interval – has been suggested as an indicator of consciousness (Allen 2013; Birch 2022). In accordance with desideratum 1b, this is supported by experiments which purport to show that humans need to perceive stimuli consciously for trace conditioning to occur (Clark and Squire 1998, 1999). This further experimental confirmation makes trace conditioning a better test of animal consciousness than other non-reflexive behaviors which cannot claim the same kind of support.[4]

As pioneered by Irvine (2020) and further developed by Mason and Lavery (2022), we can further evaluate putative tests of consciousness by searching for animals which pass the test

[4] For a discussion of trace conditioning as an indicator of animal consciousness, see Droege et al. (2021).

but are almost certainly not conscious.[5] Irvine proposes that nematode *c. elegans*, whose nervous system only comprises 302 neurons, can be assumed to lack consciousness. Intriguingly, Irvine argues that (at least simple forms of) motivational trade-off can be found in *c. elegans* which she takes as evidence that such trade-offs do not indicate consciousness. In the analysis of Mason and Lavery, multiple kinds of organism serve as test cases: plants and protozoa (P), spines disconnected from brains (S), decerebrate mammals and birds (D) and humans in unaware states, e.g. anesthesia, sleep or being exposed to subliminal stimuli (U). Given the plausible working assumption that these "S.P.U.D. subjects" are not conscious (of the relevant stimuli), putative tests of animal consciousness are ruled out if S.P.U.D. subjects pass them.

According to Mason and Lavery, this undermines (at least) unlearned affective responses like avoidance, approach, flight or vocalization as tests of consciousness. Moreover, such responses can also be modulated in S.P.U.D. subjects by affectively relevant cues and S.P.U.D. subjects can learn via Pavlovian conditioning, some forms of trace conditioning and instrumental conditioning (where innate responses to stimuli are modified via learning). Thus, collecting further empirical evidence from humans and animals lets us scrutinize our tests of consciousness themselves.

Surely, it is not impossible that S.P.U.D. subjects are conscious. Due to the diversity of views on consciousness which includes positions according to which conscious is extremely widespread (Tononi and Koch 2015) or universal (Goff 2017), we cannot find a being as an interesting test case which *everyone* agrees not to be conscious. But the belief that S.P.U.D. subjects are probably not conscious is nevertheless shared by almost all researchers which do not presuppose a particular theory of consciousness. Two points are further noteworthy: First, this method is especially credible if multiple types of S.P.U.D. subjects pass the test in question. Second, the assumption that certain S.P.U.D. subjects are not conscious is subject to revision in the light of further evidence. In particular, if it turns out that one kind of S.P.U.D. subjects can pass numerous tests of consciousness which other S.P.U.D. subjects fail, one may begin to question whether the former kind actually lacks consciousness.

[5] I don't regard it as strong evidence against a test of animal consciousness if non-conscious machines can be designed to pass the test. For if one is sufficiently ingenious, machines can be designed to pass almost any consciousness test (Shevlin 2020; Tomasik 2014). In addition, arguably tests of animal consciousness need to overcome a lower evidential bar than machine consciousness tests since the former can implicitly rely on similarity between humans and animals in virtue of common evolutionary descent which we do not share with machines and which buttresses arguments from analogy.

Crucially, it seems like it is not sufficient for applying NP that a behavior is *frequently* caused by conscious experience in humans. For NP is usually justified by a preference for more parsimonious, unified explanation. Explaining one type of behavior or capacity in recourse to two types of causes – a conscious process in humans and an unconscious process in another species – means to introduce superfluous causes, except when there is independent reason to believe that different types of causes operate. However, if there are *some* instances where the same type of behavior is caused unconsciously in humans, then the explanation via consciousness in non-human animals is not more parsimonious. If there are a conscious and an unconscious mechanism for a given behavior, we have no reason to disregard the hypothesis that the behavior is always caused unconsciously in other animals. Thus, to motivate belief in animal consciousness via NP, a behavior has to be always caused by consciousness in humans.[6]

Desideratum 1c refers to an additional improvement of the evidential strength of tests for animal consciousness. It states that the animal behavior should indicate that the animal processes stimuli in two different ways, where two types of behavioral response correspond to behavior which is caused by human conscious and unconscious processing, respectively. The recent study by Ben Haim et al. (2021) on Rhesus monkeys is unique in satisfying this desideratum.

The study employs a spatial cueing task. In the task, a target appears in one of two possible locations on a screen and the subjects have to identify the target location as quickly as possible (the subjects eye gaze is used as the mode of response). A cue precedes the target. This cue is either presented subliminally (17/33 ms) or supraliminally (250 ms). Crucially, the cue is presented in the opposite location of the target. That is, the cue predicts the appearance of the target, but it is incongruent, i.e., its location is always the opposite from the target's location.

It turns out that the pattern of results is the same for humans and monkeys: Relative to a control condition using a non-predictive cue, a supraliminally presented cue speeds up target identification (after subjects had the opportunity to learn the cue-target contingency) while a subliminally presented cue *slows down* target identification. By asking humans for verbal reports, Ben Haim et al. confirmed that humans were conscious of the supraliminally presented stimuli and did not experience the subliminal ones. Hence, the study reveals a *double*

[6] Birch (2022) allows that a behavioral indicator of consciousness may only be facilitated by consciousness, while its occurrence does not depend on consciousness. Through this facilitation, consciousness is assumed to increase the speed, reliability or a similar feature of the behavioral capacity in question. To make this claim consistent with the rationale for NP, one can point out that – through this facilitation relation – consciousness may not be necessary for a given behavioral capacity but is nevertheless necessary for performing this behavior with a certain speed, accuracy or the like. Thus, if the behavior is performed with the relevant feature, then NP can be applied.

dissociation of visual consciousness in humans, i.e., a condition in which processing of consciously accessible stimuli and of stimuli just below the threshold for conscious perception have opposite effects on performance.[7]

This is a result which goes beyond the kinds of cases merely satisfying desideratum 1b where it is only shown that particular behavior apparently cannot be performed without consciousness. For it suggests the existence of two distinct processing modes. It is not only the case that conscious perception improves performance in humans, but unconscious processing can have distinctive effects as well: it impairs performance. Since Rhesus monkeys show the same double dissociation in the same task, it seems very likely that the dissociation signifies the distinction between conscious and unconscious processing in them as well. If animals exhibit performance which is analogous to known double dissociations in humans, then this provides especially strong evidence of consciousness. Hence, up until now, the best evidence of animal consciousness we found consists in behavior which equals the behavior in double dissociations of consciousness confirmed in humans. However, the proposed account allows that weaker evidence is valuable as well. Next section, I explain and motivate three additional desiderata.

3. Theoretical integration

To draw conclusions about animal consciousness, it is common to proceed in one of two ways (Birch 2022): Either researchers rely on cognitive-behavioral evidence suggesting that animals are so complex or human-like that we should attribute consciousness to them or they infer the distribution of consciousness from a theory of the nature, function or physical substrate of consciousness which specifies necessary and sufficient conditions for being conscious (e.g., Carruthers 1998; Klein and Barron 2016; Tononi and Koch 2015). The strengths of the former approach are encapsulated in the first three desiderata.

Directly making inferences from theories of consciousness to its distribution is problematic, however, for three reasons: First, there is widespread and profound disagreement on which theory of human consciousness is true. Second, common theories of consciousness are not specific enough to have definite implications for the distribution of consciousness (Birch 2022; Shevlin 2021). Third, even if a theory of consciousness correctly describes the

[7] As pointed out by Hampton (2021), 'double dissociation' may be a misleading term since the term usually refers to procedures where two types of experimental manipulation occur and two types of behavioral outcomes are measured. This criterion is not met by Ben Haim et al. However, the fact that the experiment by Ben Haim et al. deviates from typical experiments in consciousness science (which can show double dissociations in the traditional sense) is precisely what makes it a distinctive advance in the literature.

mechanisms responsible for human consciousness, it might not apply to other animals. To pick a toy example, insects and mammals might both be able to *see*, although they see via very different types of mechanisms, e.g. eyes. Since in evolution different structures can come to fulfill similar functions, other animals might be conscious in virtue of different mechanisms.[8] This might happen either through independent origins of consciousness in other taxa or the later co-option of simple processes by more complex brain regions in mammals (Godfrey-Smith 2016b) or humans specifically.

To overcome these challenges, I aim to integrate theoretical knowledge about consciousness with empirical evidence of the types discussed last section which are to some extent neutral in respect to which theory of consciousness is true.[9] For this reason, the second category of desiderata for tests of animal consciousness refers to three features which, according to many prominent theories of consciousness, are (or indicate) crucial parts of the mechanism responsible for making a content available to conscious experience. If a test indicates the presence of a cognitive capacity which is likely a crucial component of the mechanism underlying consciousness, then this makes the test more trustworthy.

Desideratum 2a states that the test indicates the presence of selective attention in the animal in question. Selective attention is the capacity to select among all signals some for deeper, more thorough processing and amplification while diminishing or excluding others. According to standard interpretations of the global-workspace theory (Cohen and Dennett 2011; Dehaene and Naccache 2001; Mashour et al. 2020), attention is necessary for a stimulus to reach consciousness. This claim is shared by other theories of consciousness (Graziano and Webb 2015; Prinz 2012).

A tight connection between consciousness and attention can also be posited on general grounds. It has to be conceded that, as Graziano et al. (2020) point out, there are many studies showing that attention is not sufficient for consciousness. For example, in blindsight subjects a cue directing attention to the "blind" part of the visual field can improve discrimination of a stimulus without leading to conscious awareness of the stimulus (Kentridge et al. 2004). Nevertheless, to let Graziano et al. (2020) speak:

> "It is a mistake, however, to conclude that attention and awareness are independent. They are extremely difficult to separate. To hit that narrow window where the stimulus is strong enough to affect attention but not strong enough to trigger

[8] See Michel (2019) for an extensive discussion of this challenge of the multiple realizability of consciousness.
[9] This may be seen as an application of what Shevlin (2021) terms the 'modest theoretical proposal'.

> awareness, one must typically use stimuli that are masked or faint, titrated at the edge of detection."

Under normal circumstances, when we attend to a stimulus, we consciously perceive it. Besides the fact that consciousness and attention normally coincide, there are theoretical reasons to posit a close connection between the two. For they share many properties: Consciousness as well as attention are directed at a target, select particular information among a wealth of unconscious and unattended information, operate over the same information domains (for example, sensory perception, emotion and bodily feelings), influence decision-making and behavior and entail deep and thorough processing (Graziano et al. 2020). Given the intimate relation between consciousness and attention, evidence of selective attention is evidence of consciousness.

I know of no example in the literature where selective attention was tested while simultaneously satisfying desiderata 1a, 1b and 1c. However, one can just perform independent tests of the capacity for selective attention on animals which have also passed tests satisfying the preceding desiderata. Experiments on archer fish exemplify what I am inclined to count as good evidence of selective attention.

In hunting, archer fish selectively orient their body at their visual target. More compellingly, they exhibit analogous effects to selective attention in mammals. As long as the target differs in speed and direction from the "distractors", increasing the number of possible target stimuli doesn't decrease the reaction time of archer fish (Ben-Tov et al. 2015). In addition, when archer fish are shown cues that signify the probable location of an upcoming target, the reaction time tends to be lower with valid than with invalid cues. But if the target appears too long after the cue, the reversed effect sets in such that reaction times are longer for the validly cued location (Gabay et al. 2013). These findings are analogous to visual pop-out and cueing effects and inhibition of return which characterize attention and visual search in mammals (Krauzlis et al. 2018). Evidence of this type is suggestive of consciousness.

Desideratum 2b states that a test of consciousness is (ceteris paribus) better than others if it demonstrates that the animal possesses capacities for the integration of diverse and widespread information. This means that it can bring together contents from a heterogenous set of distributed cognitive systems, integrate them and transmit the results back to the cognitive systems which demand them. This sort of integration can be expected to rely on long-distance and bi-directional neural connections and to express itself in flexible actions sensitive to a multitude of contents and competing demands.

Virtually all influential theories of consciousness suggest a version of desideratum 2b. Most obviously, global transmission of information is necessary and sufficient for consciousness according to the global-workspace theory. Integrated-information and recurrent-processing theory do not require information integration to be global, but only to be recurrent. Hence, proponents of the latter two theories should agree that tests satisfying desideratum 2b support attributions of consciousness, even if failing these tests does not suggest that consciousness is absent. Other theories, like Merker's (2005) midbrain theory, link consciousness as well to some form of information integration in the service of flexible action. Finally, higher-order theorists (Gennaro 2012; Rosenthal 2005), who link consciousness with some form of metacognition, should at least claim that consciousness requires a certain level of information integration. Thus, while they won't accept passing a test satisfying 2b as compelling evidence of consciousness, they should at least grant that it is a step in the right direction.

Many different behaviors have been claimed to manifest capacities for complex information integration in animals and I cannot review them all here. I will just note two promising candidates: Conditioned place avoidance (or conditioned place preference) is the capacity to learn to avoid (prefer) *places* where one was exposed to noxious (rewarding) stimuli, like an electric shock. It was documented in fish (Dunlop et al. 2006). This form of learning places demands on information integration, since it requires connecting information about a noxious stimulus with spatial memory in the service of changing future behavior (Irvine 2020).

According to Butlin (2020), goal-directed action selection is an even better indicator of conscious experience. In goal-directed action selection, animals represent and subsequently combine information about the value of outcomes and about dependencies between possible actions and their possible outcomes. Goal-directed control contrasts with habitual control in which values are attached to the actions themselves and Pavlovian control, in which innate behavior comes to be associated with new stimuli. In contrast to these more primitive forms of action selection, goal-directed control requires making information more widely available to mechanisms involved in rational action selection (Butlin 2020, p. 120). In addition to behavioral evidence, an understanding of neural anatomy and function can contribute to illuminating to what extent information is integrated in the brain. Neural evidence has, e.g., further supported the case for bird consciousness by showing that bird brains allow higher degrees of information integration than often thought (Nieder et al. 2020; Stacho et al. 2020).[10]

[10] Cross-modal learning is another potential piece of evidence for the integration of relevant mental contents (Mudrik et al. 2014).

Perhaps these two forms of integration are indeed sufficient for consciousness. In any case, their presence should make us more confident that a given animal is conscious. The same holds for desideratum 2c according to which, ceteris paribus, passing a test of consciousness is a stronger sign of consciousness if it presupposes metacognitive abilities. In adult neurotypical humans, consciousness and metacognition normally coincide. If we consciously perceive something, then we know that we perceive it. This suggests that metacognition may be associated with consciousness in other animals too.

Furthermore, inspired in part by this commonplace, higher-order theories of consciousness argue that metacognition is necessary for consciousness and a certain kind of metacognition (for instance, non-inferentially acquired higher-order thought) is sufficient for consciousness. Other theories, like global-workspace theory, are consistent with the claim that metacognition requires consciousness. For contents arguably need to be made globally available to be represented by further cognitive states.

A putative way to examine metacognitive abilities is uncertainty monitoring. Uncertainty monitoring is tested by offering animals the chance to opt out of difficult discrimination tasks to avoid punishment while also foregoing potential reward. If an animal opts out more on difficult than easy tasks and its average performance on difficult tasks improves in response, then it is said to track its own uncertainty regarding the correct decision. The capacity to monitor uncertainty of one's own mental states might qualify as a form of metacognition. While it has been disputed that tests of uncertainty monitoring actually track metacognitive abilities (Carruthers and Williams 2019; Le Pelley 2012),[11] even many skeptical views might be compatible with allowing that these tests demonstrate a version of implicit or procedural metacognition (Proust 2019). Whether the relevant form of metacognition is sufficient for consciousness depends on the theory of consciousness, including varieties of higher-order theories, under consideration. Different variants of the higher-order theory place different demands on the form of metacognition required for consciousness (Gennaro 2012; Lau 2022; Rosenthal 2005). Nevertheless, a wide array of theories of consciousness implies that evidence of some form of metacognition is relevant evidence of consciousness. It would be of considerable importance if uncertainty monitoring manifests a sufficiently sophisticated form of metacognition to satisfy desideratum 2c, since uncertainty monitoring has been claimed to be exhibited by – among other species – rats and bees (Perry and Barron 2013).

[11] For instance, Le Pelley argues that simple associative learning can explain the same results. By contrast, Carruthers and Williams hold that first-order estimates of risk are sufficient to pass tests for uncertainty monitoring.

To summarize, our general conception of how consciousness works – in conjunction with an overlapping consensus of many prominent theories of consciousness – suggests that tests indicating the presence of selective attention, widespread information integration and metacognition justify increases in our confidence that animals passing them are conscious. Before moving to the next section, I make explicit that the desiderata proposed here also rely on some sort of analogy argument. Namely, they rely on the assumption that human and non-human animal consciousness are in some respect similar. This similarity is presupposed because our general conception of consciousness as well as our theories of consciousness are mainly derived from investigating what consciousness is in humans.

However, if one takes these desiderata only as virtues of positive tests of the *presence* (as opposed to the absence) of animal consciousness, then the relevant assumption of similarity is eminently plausible. At some point, if animals display enough of the mechanisms which are candidates for contributing to consciousness in humans, we should count them as conscious too. It is not the case that mechanisms which are sufficient for consciousness in humans could plausibly be seen as insufficient for consciousness in non-human animals.

Before moving on to the remaining desiderata, I note that the rationale for including the three desiderata of this section suggests that the list presented here needs to be updated in line with scientific progress. If our theoretical knowledge of consciousness evolves, then the desiderata motivated by theories of consciousness might change as well.

4. The icing on the cake

In this section, I will describe the remaining two desiderata. Both are not self-standing. However, tests of consciousness which already satisfy other desiderata can be significantly improved by modifying them to include these two desiderata as well. According to desideratum 3a, tests which satisfy desideratum 1a, 1b or 1c are improved when they involve differential responses to the *same* type of stimulus. To make this concrete: The animal studies on trace conditioning and double dissociations elicited by spatial priming, which we have discussed in section 2, use either contrasting pairs of stimuli or just one type of stimulus. In the study of Ben Haim et al., supraliminal and subliminal stimuli are presented with different durations. This is a vulnerability since it opens up space for the criticism that the double dissociation in performance in the supra- vs. subliminal condition might be due not to a difference in consciousness, but in signal strength (Crump and Birch 2021). Perhaps it is just easier to learn the cue-target contingency in the supraliminal condition since stronger signals facilitate

learning.[12] On this interpretation, the study does not succeed in showing that there are two distinct processing modes.

Available studies of trace conditioning in animals don't contrast a supraliminal with a subliminal condition at all. They merely demonstrate that the animal, in general, is able to perform trace conditioning. For this reason, one may doubt as well that the capacity for trace conditioning depends on consciousness in the animal in question, even if this is the case in humans.

The appropriate response to both objections is to employ experimental paradigms – well-established from human consciousness science – in which the conscious experience of a stimulus varies while objective properties of the stimulus remain as constant as possible. One example is backward masking, another is binocular rivalry. Let's first focus on the former.[13] In backward masking, a second stimulus (the 'mask') is presented shortly after a first stimulus. The first stimulus does not reach consciousness because it was followed by the mask, i.e., the first stimulus would have been seen consciously, if the second stimulus hadn't occurred. Backward masking is a valuable paradigm in consciousness science because it is a targeted experimental manipulation of conscious perception which leaves the stimulus and the situation relatively unchanged. That is, if a certain effect occurs only in respect to unmasked stimuli, not in response to masked ones, this difference is less likely to be caused by a stimulus difference and is more likely caused by the difference in consciousness.

Let's apply this to our examples. In the study by Ben Haim et al., one should present the two cues with the same duration but use a masked and an unmasked condition. In addition, one should test whether animals are only able to perform trace conditioning in respect to unmasked stimuli. If the results of the study by Ben Haim et al. are the same when they employ masking instead of varying stimulus durations and if trace conditioning requires unmasked stimuli, this supports the contention that these tests indeed track conscious experience.

However, the masked and the unmasked condition are not *identical* in the physical properties of the stimulus. After all, the masked condition contains an additional stimulus acting as the mask. One might be worried that even this minor stimulus difference may confound

[12] Ben Haim et al. manage to rule out this hypothesis for humans by performing a modified version of the experiment, in which they informed the human subjects that there are subliminal cues. In this condition, human performance in identifying the target location when presented the subliminal cue improves relative to the original experiment. Since the signal strength is constant across both conditions, this performance enhancement can be traced to the fact that subjects manage to become conscious of more of the subliminal stimuli. However, since this strategy relies on verbal report, it is not available in respect to non-human animals.
[13] In respect to trace conditioning and other learning capacities, the use of backward masking has been advocated by Birch (Birch 2022).

studies on consciousness (Lau 2022, chapter 2), including animal consciousness. For this reason, further techniques from human consciousness science should also be considered.

In binocular rivalry paradigms, different images are presented separately and simultaneously to each eye of the subject. This causes an alternation of the consciously perceived image. Sometimes the image presented to the left eye is conscious, sometimes the stimulus presented to the right eye is seen consciously instead. This alternation of consciousness happens *without* any change in the physical properties of the stimulus. This allows a comparison of conscious and unconscious conditions which is not confounded by the physical properties of the stimuli (Lau 2022, chapter 2). While adapting binocular rivalry paradigms to non-human animals is challenging, it has been done with monkeys (Logothetis and Schall 1990). In addition, other cases of multi-stable perception, some belonging to other sensory modalities, can be used (Schwartz et al. 2012).

There is no principled reason why paradigms like masking and binocular rivalry should not be transferable to non-human animals, including invertebrates. I do not want to minimize the methodological challenges involved in applying masking paradigms to other mammals, let alone invertebrates. However, such paradigms have at least been transferred to rats and it seems that masking effects indeed occur in rats (Dell et al. 2019). In addition, analogous masking effects are also found in other sensory modalities, e.g. audition (Oxenham 2013), which opens the possibility to apply masking to a wider range of species. While the jury is still out, there is currently no reason to deny that binocular rivalry, masking or some similar effects can occur in most non-human animals.

Desideratum 3b states that tests of consciousness should not be considered in isolation, but that researchers should examine correlations between different capacities which are claimed to indicate consciousness.[14] If consciousness is responsible for the capacity to pass different sorts of tests, then capacities to pass these tests likely correlate with each other. This correlation can either be tested by looking at different animal species or at different experimental conditions applied to the same species. That is, we should expect that an animal species tends to pass many tests of consciousness if it passes a single one. Likewise, we should expect that several cognitive capacities which are posited to require consciousness are turned on and off in the same situations. If a stimulus is perceived consciously, many consciousness-dependent capacities should be able to operate on it. If it is unconscious, then they should be unavailable.

[14] This desideratum is heavily inspired by the 'theory-light' approach for investigating animal consciousness championed by Birch (Birch 2022).

In total, this means that consciousness-dependent cognitive capacities should also correlate over different stimulus conditions. For instance, if a stimulus is masked, a significant fraction of them should be turned off (Birch 2022). If those correlations between different tests are found, the best explanation of them seems to be that they all depend on a shared cognitive process: consciousness.[15] Note, however, that the possession of consciousness does not require animals to exhibit the relevant correlations since our desiderata are primarily intended to be desiderata for positive tests of the presence, rather than the absence, of consciousness.

Let us summarize what we have learned from all desiderata. A perfect set of empirical tests of the presence of animal consciousness in an animal species S is a set which possesses several features: It reveals whether S possesses several behavioral capacities which are non-reflexive and – as supported by experimental evidence – seem to require consciousness in humans. Some of these capacities involve two distinct processing modes, whose effects tend to correspond to the effects of human conscious and unconscious processing, respectively. In addition, the tests examine whether S is capable of selective attention, widespread and complex integration of information and metacognition. Moreover, the tests apply paradigms from human consciousness science which allow for specific comparisons between conscious and unconscious perception whenever useful. Finally, the application of this set of tests includes an examination of whether capacities to pass various tests of consciousness correlate over different species and over different task conditions.

This set of tests is desirable because it balances two competing demands: It is *informative* in that it can potentially be applied to many different kinds of species, including non-mammals and even invertebrates. Also, it is an open empirical question to what extent different species can pass these tests. At the same time, passing many of the tests contained within this set provides strong evidence that an animal is conscious. While a resolute skeptic may always insist that animals may still lack consciousness, even if their behavior and cognitive processing is similar to human consciousness-relevant capacities to a large extent and regarding many details, this stance successively loses plausibility the more pronounced the similarities

[15] This method presupposes that consciousness facilitates the same or similar clusters of cognitive abilities in humans and non-human animals. This is most likely if the function of consciousness was constant over long periods of evolutionary history. One might consider this assumption of evolutionary constancy implausible. The advantage of this method is, however, that the use of this method is a way to test the assumption of evolutionary constancy. If clusters of abilities which indicate consciousness in humans are found, then this supports the assumption that consciousness' function is relatively constant across species. If such clusters are not found, then this route of providing evidence of animal consciousness is blocked. However, not finding such a cluster does not significantly count against attributions of consciousness, as the desiderata concern only positive tests of consciousness.

between human and animal cognition turn out to be. If an animal passes the whole set of tests I proposed with flying colors, we should strongly prefer the hypothesis that the animal is conscious to its negation.

The list of desiderata presented here is chiefly based on cognitive features which are instances of general cognitive sophistication: flexible behavior, selective attention, information integration and metacognition. However, there are many theoretical conceptions of consciousness according to which the basic ability to feel does not require a particularly large amount of cognitive complexity (Godfrey-Smith 2020; Merker 2005; Solms 2021). My approach is compatible with these conceptions since it only describes desiderata for positive, not negative, tests of animal consciousness. While the list proposed here captures features of valuable positive tests of animal consciousness, the view that consciousness does not presuppose particularly large cognitive complexity would suggest that alternative, less demanding lists of desiderata can be designed as well. Yet, in the absence of knowledge regarding the correct theory of consciousness, the more ecumenical approach developed here has an advantage.

In the remainder of the paper, I will use this list of desiderata to evaluate a proposed test and putative criteria of animal (pain) consciousness from the literature. This is intended to illustrate the practical value of these desiderata. It will be shown that the desiderata illuminate the virtues and shortcomings of this test and localize the need to supplant it by complementary indicators.

5. Motivational trade-off and criteria for animal sentience

My example of the application of the desiderata is trade-off behavior in response to noxious stimuli which has been taken as evidence of consciousness by many authors (e.g., Sneddon et al. 2014; Tye 2017). Motivational trade-off behavior has been found in fish (Millsopp and Laming 2008), hermit crabs (Appel and Elwood 2009) and bees (Gibbons et al. 2022). In the experiment of Millsopp and Laming (2008), goldfish reduced their feeding attempts in a part of an aquarium where they received a shock. The more intensive the shocks were, the less feeding attempts were made by the fish. However, when the fish were increasingly food-deprived, the number and duration of feeding attempts increased. Consequently, fish seem to trade off their need for food with their aversion to noxious stimuli. This would entail that they integrate different kinds of information and exploit this in the service of flexible and appropriate action.

How does this putative indicator of consciousness fare in respect to our desiderata? Behavior shown in this experiment is not a simple reflex (desideratum 1a), doesn't involve metacognition (2c) or double dissociations of consciousness (1c) and cannot be refined by masking in any obvious way (because neither hunger nor electrical shocks can be masked) (3a). While someone may argue that trade-offs presuppose selective attention (2a), this connection is at least unclear which is why they cannot serve as an indicator for selective attention. Hence, this test clearly satisfies desideratum 1a while it does not satisfy 1c, 2a, 2c and 3a. There is no direct evidence on the relation between human trade-off behavior and consciousness, thus 1b is not fulfilled either. Irvine (2020) even claims that *C. elegans* performs trade-offs of the relevant kind which would undermine them as a test of consciousness.

The allure of motivational trade-offs as a test of consciousness stems mainly from desideratum 2b. Trade-offs seem to require some sophisticated form of information integration. If this is true, then they can serve as part of our ideal set of consciousness tests, despite not fulfilling other desiderata. The list of desiderata points to the requirements for making trade-off paradigms a valuable consciousness test. The notion of motivational trade-off has to be made more precise such that it fulfils three requirements: First, performing trade-offs requires integration of contents from numerous cognitive systems. Second, trade-offs can be shown to require consciousness in humans. Third, *C. elegans* cannot perform trade-offs (in the relevant sense). If those conditions are met, motivational trade-offs can serve as an important test of consciousness. Crucially, to increase our confidence that the species which performs trade-offs is conscious further, we would need to address the desiderata which are not satisfied by this test.

A recent experiment in bumblebees by Gibbons et al. (2022) incorporates an improvement of motivational trade-off paradigms which fits with some of the suggestions made here. In this study, the trade-off (between noxious heat and a rewarding sucrose solution) was supplemented by a memory component. The bees could not detect the concentration of sucrose without feeding on it (since it doesn't smell) and the number of landings on the feeders ("testing out") decreased over the course of the experiment. For this reason, since trade-off behavior was indeed found, the bees must have learned to associate the content of the feeder with its spatial location or a color cue (which has been presented). It follows that "the trade-off relied on associative memories, rather than direct experience, of the stimuli" (Gibbons et al. 2022). Since these trade-offs depend on associative memory, their demand for information integration is increased. Moreover, there is no evidence that trade-offs which rely on associative memory can be performed by *C. elegans* (although there is also no evidence to the contrary).

In this context, it is instructive to further look at the list of proposed indicators of animal pain experience provided by Crump et al. (2022). They propose eight criteria for sentience, focusing on pain experience: 1. Nociception. 2. Sensory integration. 3. Integrated nociception. 4. Analgesia. 5. Motivational trade-offs. 6. Flexible self-protection. 7. Associative Learning. 8. Analgesia preference.

How can we evaluate this list, based on our list of desiderata? We already discussed motivational trade-offs. Notably, criterion 2 and 3 have a similar rationale as motivational trade-off paradigms: they are likewise based on the idea that information integration is evidence of consciousness. Since these criteria are not connected to further evidence regarding the specific degree of integration sufficient for consciousness, criterion 2, 3 and 5 stand and fall together: if it turns out that an animal can satisfy one of these criteria without being conscious, then there is reason to think that it might satisfy all three of them without being conscious.

Nociception (criterion 1) is plausibly necessary for pain processing but not relevant to consciousness as such. Similarly, the fact that an animal responds to analgesics and possesses an endogenous transmitter system modulating its response to noxious stimuli (criterion 4) tells us something about its nociceptive system but is no evidence that it experiences nociception consciously. None of the other criteria satisfies more than desideratum 1a, i.e., displaying non-reflexive human-like responses to stimuli: flexible self-protection, associative learning and preference for and self-administration of analgesics are all forms of behavior which we find in humans and which are not simple reflexes. However, they satisfy no other desiderata. In respect to associative learning, there is even evidence that it can occur unconsciously (Mason and Lavery 2022).

This assessment suggests that the criteria for animal sentience championed by Crump et al. encapsulate rather weak evidence when compared with the list of desiderata proposed here. This traces back to their difference in function: While the list of desiderata proposed here aims to inspire and guide future research, the criteria by Crump et al. are used to make policy-relevant assessments of sentience in the present. Thus, they are limited to criteria where there already are many useful experimental studies available.

While I agree with Crump et al. that their criteria are suitable to determine whether animals fall within the scope of a precautionary principle such that they should be protected by animal welfare law, their criteria provide rather weak evidence for the purposes of animal consciousness science. In particular, their framework could be fruitfully amended by adding criteria which are not based on evidence of information integration and merely non-reflexive responses to noxious stimuli.

This illustrates that the desiderata of consciousness we have converged on can be used to assess the evidential profile and overall strength of both particular tests of consciousness and of sets of tests of consciousness. Also, this list reveals shortcomings and methods of improvement and points to relevant open question. Sometimes the correct response will be to improve the test in question and collect further support for the claim that it satisfies the respective desideratum. In other cases, it will be best to replace the test or complement it with further tests which tackle other desiderata.

6. Conclusion

I have proposed a list of eight desiderata, divided into three categories, which can be used to assess the evidential value of putative tests of animal consciousness and suggest further improvements. The first category comprises desiderata which strengthen inferences that are based on analogies between human and animal behavior. Non-reflexive behavior which seems to depend on consciousness in humans and which ideally is symptomatic of a double dissociation of consciousness should be the target of tests of animal consciousness. The second category comprises desiderata derived from our theoretical understanding of consciousness. Tests which indicate the presence of selective attention, widespread information integration and metacognition support attributions of consciousness. Finally, according to the third category, tests which satisfy some of the other desiderata can be enhanced by employing paradigms from human consciousness science like masking or multi-stable perception and by discovering correlations between different cognitive capacities which are all individually thought to be indicative of consciousness.

Thus, the preceding discussion has clarified that evidence in virtue of analogy comes in different degrees. It starts with loose analogies (desideratum 1a) and then moves to analogies which are closer (1c, 3b) and more firmly grounded in human consciousness science (1b, 3a). Moreover, it has been shown that several indicators of consciousness (selective attention, information integration and metacognition) are relatively robust to variation in which theory of consciousness is true. Finally, the framework emphasizes the need to systematically evaluate sets of tests of consciousness and their respective focus and limitations, not just individual tests. Most of these desiderata have already inspired some isolated research projects on animal consciousness but there is much to be gained from systematizing, justifying and combining them.

References

Allen, C. (2013). Fish Cognition and Consciousness. *Journal of Agricultural and Environmental Ethics*, *26*(1), 25–39. https://doi.org/10.1007/s10806-011-9364-9

Appel, M., & Elwood, R. W. (2009). Motivational trade-offs and potential pain experience in hermit crabs. *Applied Animal Behaviour Science*, *119*(1), 120–124. https://doi.org/10.1016/j.applanim.2009.03.013

Ben-Haim, M. S., Dal Monte, O., Fagan, N. A., Dunham, Y., Hassin, R. R., Chang, S. W. C., & Santos, L. R. (2021). Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (Macaca mulatta). *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.2017543118

Ben-Tov, M., Donchin, O., Ben-Shahar, O., & Segev, R. (2015). Pop-out in visual search of moving targets in the archer fish. *Nature Communications*, *6*(1). https://doi.org/10.1038/ncomms7476

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

Brown, C. (2015). Fish intelligence, sentience and ethics. *Animal Cognition*, *18*(1), 1–17. https://doi.org/10.1007/s10071-014-0761-0

Browning, H., & Veit, W. (2020). The Measurement Problem of Consciousness. *Philosophical Topics, 48*(1), 85–108. https://doi.org/10.5840/philtopics20204815

Butlin, P. (2020). Affective Experience and Evidence for Animal Consciousness. *Philosophical Topics*, *48*(1), 109–127. https://doi.org/10.5840/philtopics20204816

Carruthers, P. (1998). Natural theories of consciousness. *European Journal of Philosophy*, *6*(2), 203–22.

Carruthers, P. (2020). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford, New York: Oxford University Press.

Carruthers, P., & Williams, D. M. (2019). Comparative metacognition. *Animal Behavior and Cognition*, *6*(4), 278–288. https://doi.org/10.26451/abc.06.04.08.2019

Clark, R. E., & Squire, L. R. (1998). Classical Conditioning and Brain Systems: The Role of Awareness. *Science*, *280*(5360), 77–81. https://doi.org/10.1126/science.280.5360.77

Clark, R. E., & Squire, L. R. (1999). Human Eyeblink Classical Conditioning: Effects of Manipulating Awareness of the Stimulus Contingencies. *Psychological Science*, *10*(1), 14–18. https://doi.org/10.1111/1467-9280.00099

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358--364.

Crump, A., & Birch, J. (2021). Separating Conscious and Unconscious Perception in Animals. *Learning and Behavior*, *49*(4).

Crump, A., & Birch, J. (2022). Animal Consciousness: The Interplay of Neural and Behavioural Evidence. *Journal of Consciousness Studies*, *29*(3–4), 104–128.

Crump, A., Browning, H., Schnell, A., Burn, C., & Birch, J. (2022). Sentience in decapod crustaceans: A general framework and review of the evidence. *Animal Sentience*, *7*(32). https://doi.org/10.51291/2377-7478.1691

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1), 1–37.

Dell, K. L., Arabzadeh, E., & Price, N. S. C. (2019). Differences in perceptual masking between humans and rats. *Brain and Behavior*, *9*(9). https://doi.org/10.1002/brb3.1368

Dimitrijević, M. R., & Nathan, P. W. (1968). Studies of spasticity in man: Analysis of reflex activity evoked by noxious cutaneous stimulation. *Brain*, *91*(2), 349–368. https://doi.org/10.1093/brain/91.2.349

Droege, P., Weiss, D. J., Schwob, N., & Braithwaite, V. (2021). Trace conditioning as a test for animal consciousness: a new approach. *Animal Cognition*, *24*(6), 1299–1304.

https://doi.org/10.1007/s10071-021-01522-3

Dunlop, R., Millsopp, S., & Laming, P. (2006). Avoidance learning in goldfish (Carassius auratus) and trout (Oncorhynchus mykiss) and implications for pain perception. *Applied Animal Behaviour Science*, *97*(2–4), 255–271. https://doi.org/10.1016/j.applanim.2005.06.018

Gabay, S., Leibovich, T., Ben-Simon, A., Henik, A., & Segev, R. (2013). Inhibition of return in the archer fish. *Nature Communications*, *4*(1), 1657. https://doi.org/10.1038/ncomms2644

Gennaro, R. J. (2012). *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*. MIT Press.

Gibbons, M., Versace, E., Crump, A., Baran, B., & Chittka, L. (2022). Motivational trade-offs and modulation of nociception in bumblebees. *Proceedings of the National Academy of Sciences*, *119*(31), e2205821119. https://doi.org/10.1073/pnas.2205821119

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. The MIT Press. https://doi.org/10.7551/mitpress/11006.001.0001

Godfrey-Smith, P. (2016a). *Other minds: the octopus, The sea, and the deep origins of consciousness*. Farrar, Strauss and Giroux.

Godfrey-Smith, P. (2016b). Mind, Matter, and Metabolism. *Journal of Philosophy*, *113*(10), 481–506. https://doi.org/10.5840/jphil20161131034

Godfrey-Smith, P. (2020). *Metazoa: Animal minds and the birth of consciousness*. William Collins.

Goff, P. (2017). *Consciousness and Fundamental Reality* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190677015.001.0001

Graziano, M. S. A., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2020). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*, *37*(3–4), 155–172. https://doi.org/10.1080/02643294.2019.1670630

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00500

Hampton, R. R. (2021). Animal consciousness: Should a new behavioral correlate in monkeys persuade agnostics? *Current Biology*, *31*(12), R801–R803. https://doi.org/10.1016/j.cub.2021.05.020

Hoffman, K. N. (2020). Subjective Experience in Explanations of Animal PTSD Behavior. *Philosophical Topics*, *48*(1), 155–175. https://doi.org/10.5840/philtopics20204818

Irvine, E. (2020). Developing Valid Behavioral Indicators of Animal Pain. *Philosophical Topics*, *48*(1), 129–153. https://doi.org/10.5840/philtopics20204817

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

Key, B. (2016). Why fish do not feel pain. *Animal Sentience*, *1*(3). https://doi.org/10.51291/2377-7478.1011

Klein, C., & Barron, A. B. (2016). Insects have the capacity for subjective experience. *Animal Sentience*, *1*(9). https://doi.org/10.51291/2377-7478.1113

Krauzlis, R. J., Bogadhi, A. R., Herman, J. P., & Bollimunta, A. (2018). Selective attention without a neocortex. *Cortex*, *102*, 161–175. https://doi.org/10.1016/j.cortex.2017.08.026

Lau, H. (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford, New York: Oxford University Press.

Laureys, S. (2007). Eyes Open, Brain Shut. *Scientific American*, *296*(5), 84–89. https://doi.org/10.1038/scientificamerican0507-84

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 686–708. https://doi.org/10.1037/a0026478

Logothetis, N. K., & Schall, J. D. (1990). Binocular motion rivalry in macaque monkeys: Eye dominance and tracking eye movements. *Vision Research*, *30*(10), 1409–1419. https://doi.org/10.1016/0042-6989(90)90022-D

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

Mason, G. J., & Lavery, J. M. (2022). What Is It Like to Be a Bass? Red Herrings, Fish Pain and the Study of Animal Sentience. *Frontiers in Veterinary Science*, *9*. https://www.frontiersin.org/articles/10.3389/fvets.2022.788289. Accessed 5 July 2022

Merker, B. (2005). The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, *14*(1), 89–114. https://doi.org/10.1016/S1053-8100(03)00002-3

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

Millsopp, S., & Laming, P. (2008). Trade-offs between feeding and shock avoidance in goldfish (Carassius auratus). *Applied Animal Behaviour Science*, *113*(1–3), 247–254. https://doi.org/10.1016/j.applanim.2007.11.004

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, *18*(9), 488–496. https://doi.org/10.1016/j.tics.2014.04.009

Nieder, A., Wagener, L., & Rinnert, P. (2020). A neural correlate of sensory consciousness in a corvid bird. *Science*, *369*(6511), 1626–1629. https://doi.org/10.1126/science.abb1447

Oxenham, A. J. (2013). Mechanisms and mechanics of auditory masking. *The Journal of Physiology*, *591*(Pt 10), 2375. https://doi.org/10.1113/jphysiol.2013.254490

Perry, C. J., & Barron, A. B. (2013). Honey bees selectively avoid difficult choices. *Proceedings of the National Academy of Sciences*, *110*(47), 19155–19159. https://doi.org/10.1073/pnas.1314571110

Prinz, J. (2012). *The Conscious Brain: How Attention Engenders Experience*. Oxford University Press USA.

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.

Proust, J. P. (2019). From comparative studies to interdisciplinary research on metacognition. *Animal Behavior and Cognition*, *6*(4), 309–328. https://doi.org/10.26451/abc.06.04.10.2019

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

Schwartz, J.-L., Grimault, N., Hupé, J.-M., Moore, B. C. J., & Pressnitzer, D. (2012). Multistability in perception: binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1591), 896–905. https://doi.org/10.1098/rstb.2011.0254

Shea, N., & Bayne, T. (2010). The Vegetative State and the Science of Consciousness. *The British Journal for the Philosophy of Science*, *61*(3), 459–484. https://doi.org/10.1093/bjps/axp046

Shea, N., & Heyes, C. (2010). Metamemory as evidence of animal consciousness: the type that does the trick. *Biology & Philosophy*, *25*(1), 95–110. https://doi.org/10.1007/s10539-009-9171-0

Shevlin, H. (2020). General intelligence: an ecumenical heuristic for artificial consciousness research? *Journal of Artificial Intelligence and Consciousness*. https://doi.org/10.17863/CAM.52059

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

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour*, *97*, 201–212. https://doi.org/10.1016/j.anbehav.2014.09.007

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. W. W. Norton & Co.

Stacho, M., Herold, C., Rook, N., Wagner, H., Axer, M., Amunts, K., & Güntürkün, O. (2020). A cortex-like canonical circuit in the avian forebrain. *Science*, *369*(6511), eabc5534.

https://doi.org/10.1126/science.abc5534

Tomasik, B. (2014). Do Artificial Reinforcement-Learning Agents Matter Morally? *arXiv:1410.8233 [cs]*. http://arxiv.org/abs/1410.8233. Accessed 14 November 2021

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

Tye, M. (2017). *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190278014.001.0001

Yuki, S., & Okanoya, K. (2017). Rats show adaptive choice in a metacognitive task with high uncertainty. *Journal of Experimental Psychology: Animal Learning and Cognition*, *43*(1), 109–118. https://doi.org/10.1037/xan0000130