Please cite as: Browning, H. & Veit, W. (2021). The Measurement Problem of Consciousness. Preprint.

Check https://walterveit.com/publications/ for citation details once published

The Measurement Problem of Consciousness

Heather Browning & Walter Veit

Abstract

This paper addresses what we consider to be the most pressing challenge for the emerging science of consciousness: the measurement problem of consciousness. That is, by what methods can we determine the presence of and properties of consciousness? Most methods are currently developed through evaluation of the presence of consciousness in humans and here we argue that there are particular problems in application of these methods to non-human cases - what we call the indicator validity problem and the extrapolation problem. The first is a problem with the application of indicators developed using the differences in conscious and unconscious processing in humans to the identification of other conscious vs. non-conscious organisms or systems. The second is a problem in extrapolating any indicators developed in humans or other organisms to artificial systems. However, while pressing ethical concerns add urgency in the attribution of consciousness and its attendant moral status to non-human animals and intelligent machines, we cannot wait for certainty and we advocate the use of a precautionary principle to avoid doing unintentional harm. We also intend that the considerations and limitations discussed in this paper can be used to further analyse and refine the methods of consciousness science with the hope that one day we may be able to solve the measurement problem of consciousness.

Keywords: consciousness; measurement; artificial intelligence; animal minds

1 Introduction

In recent decades, the study of consciousness has been undergoing a transition from a purely philosophical mode of investigation towards a genuine science of consciousness, thus giving rise to the *measurement problem of consciousness* (henceforth: **MPC**). The problem can be summarized as follows: by what methods can we determine the presence and properties of consciousness? In our view, the **MPC** is the biggest challenge for the study of consciousness. Compared to some of the other problems recognized in the literature, however, this problem has received surprisingly little attention from the public and scientists alike. Ability to measure is often seen merely as a means to comparing different theories and their progress, rather than as a deeper philosophical problem as we will argue here.[1] The **MPC** is not, however, only a scientific problem but also a pressing moral and political one because the presence of consciousness is seen by many as a necessary condition for the attribution of moral status (Singer 1975; Browning 2019a,2020a; Mellor and Beausoleil 2015; Duncan 2002).

Here, we will look at three distinct, yet closely related problems within and relating to the **MPC**: i) the *indicator validity problem*, (ii) the *extrapolation problem*, and (iii) the *moral problem*. The first concerns the problem of validating indicators of an inaccessible target such as conscious experience - particularly in the distinction between unconscious and non-conscious states, one we argue has previously been overlooked in discussions of this problem. The second is a different type of problem, that of inferring the usefulness of measures developed in humans to non-humans or artificial systems.[2] Both problems pose their strongest challenge for the attribution of consciousness to machines and those non-human animals more distantly related to humans. We will argue that the indicator validity problem should make us more cautious about our ability to apply indicators developed using measures of consciousness and unconsciousness in humans to attempts to identify the presence (or lack) of consciousness in other entities, such as non-human animals or machines. We will then show that although there are possible solutions available in the animal case, these will not apply in the machine case. This should lower our confidence in extrapolating current indicators to determine whether machines possess consciousness. Not only do these cases hold additional measurement challenges, as we cannot rely on the subjective measures used in humans (i.e. self-report), but the subjects of these projects are also the primary beneficiaries of a solution in terms of receiving moral consideration, i.e. the moral problem (iii). This, while not directly a

[1] Cf. Sandberg et al. 2010 and Seth et al. 2008.

[2] This problem can also be found (albeit to a lesser degree) when measures are extrapolated across humans.

measurement problem, concerns the moral and ethical implications of the **MPC** itself. A solution to the **MPC** is thus crucial not only for consciousness scientists but also philosophers, ethicists, and public policy-makers alike.

Before we begin, it will prove useful to clarify the goals and scope of this article. Firstly, it is not our primary aim here to provide a solution to the **MPC**. Indeed, it would be misleading to suggest that any single paper could tackle the task of slaying this many-headed conceptual and empirical hydra. Rather, this paper is intended to elucidate some of the particular ways in which this problem manifests itself in cases of non-human consciousness. It is our goal to clarify these subtleties and partially untangle this muddled debate. The paper is structured as follows. In Section 2, we provide a short review of the extant literature on the measurement of consciousness and offer a more detailed analysis of the **MPC**. In Sections 3 and 4, we describe the two primary strands of the **MPC** that apply to measurement of non-human consciousness: the indicator validity problem and the extrapolation problem. In Section 5, we discuss the options for proceeding if we are unable to solve the **MPC** in non-human cases, endorsing a version of the precautionary principle in order to address the current ethical challenges arising from nonhuman animals and intelligent machines. Finally, in Section 6, we conclude the discussion and provide some suggestions in light of our arguments as to how we might look to overcome the **MPC** in future research.

2 Measuring Consciousness

The tradition of behaviourism in the early 20th century attempted to banish any and all mentalistic concepts from psychology (Skinner 1953; Watson 1913), replacing them with behavioural descriptions. While these ideas have subsequently been largely overturned, their shadow still looms large over the science of consciousness. In the continuing spirit of behaviourism, the study of consciousness is still often seen as something less rigorous and scientific than other fields in the cognitive sciences, due to the subjective nature of its subject matter.

Researchers working on consciousness report more difficulty in receiving funding for their consciousness-related projects and a harder time on the job market than their colleagues, though this effect seems to be more pronounced among neuroscientists than philosophers (Michel et al. 2018). This cannot solely be explained by a negative perception of the rigour of consciousness science, as while neuroscientists themselves have indicated that they believe the science of consciousness to be less rigorous than other fields within neuroscience, they also report social neuroscience to be even less so, yet more likely to receive jobs and funding in this area (Michel et al. 2018). The reason for this lower status awarded to

consciousness studies is thus more likely conceptual in nature, relating to the particular ways in which we understand consciousness; a fact which deserves proper philosophical attention. Though there is limited space here to do justice to this problem, we will sketch out some potential reasons for it below.

Partially, these results can be explained by the existence of what Van Gulick (2018) has dubbed the three 'problems of consciousness' - the 'what', 'how' and 'why' questions. The 'what' question is a descriptive question, asking what consciousness is and what its principal features are. The 'how' question is an explanatory question, seeking to find how consciousness comes to exist; in particular how consciousness might arise from non-conscious entities. The 'why' question is a functional question, looking at why consciousness exists: whether it has a function, what this function might be, and whether it can even play a causal role. To these we add a fourth, and arguably more fundamental, problem: *the measurement problem of consciousness*. Rather than a question about the features of consciousness itself, this is a question about how we can come to identify the presence of consciousness, or know of its properties. There is, of course, overlap between the **MPC** and the other three problems, as there are also interactions between these listed problems. Where consciousness comes from, why it exists, and what role it plays are important questions, but none can be answered on their own. The answers to each of these questions will depend on and inform one another. Yet, although they are interconnected, they are still conceptually distinct. So too for *the measurement problem of consciousness*. Indeed, it is only by making progress on the **MPC** and thus increasing our confidence in our ability to measure consciousness that we may be able to gain empirical traction on the other questions.

The **MPC** is primarily concerned with how we might identify the presence (or absence) of consciousness, determine the degree to which it is present[3], and identify some of its features. The problem of consciousness is its subjective nature. Consciousness itself is inaccessible to direct measurement, as conscious experience is subjective and its contents only accessible to the subject themselves (Nagel 1974). Instead, its measurement must rely on observation of indirect effects, such as changes in physiology or behaviour. In the case of humans, researchers often rely on the reports of study-subjects (Michel 2017), though these reports may be unreliable, for reasons discussed in Section 3. These subjective methods are also not applicable to animals, who are unable to verbally report on their experiences.[4] They may be applicable to machines, but only where machines are programmed to accurately

[3] It is contentious whether or not consciousness comes in degrees (Shevlin 2021; Rosenthal 2019; Papineau 2003); we do not take a stance here, as whether or not consciousness comes in degrees doesn't make a difference for the arguments we present in this paper.

[4] Though some animals may be trained to 'report' behaviourally, at least at a very basic level, through pressing buttons or levers (Ginsburg and Jablonka 2019).

provide such reports. While some might therefore see the entire program as a dead-end, due to the necessarily private subjective nature of conscious experience, an informal survey indicated that about 78% (of both experts and non-experts) are under the impression that there is progress in the field (Michel et al. 2018). In spite of scientific progress in our understanding (Block et al. 2014), there is still a complete lack of consensus as to how consciousness can and should be measured.

Part of the problem is the large number of different competing theories on the nature of consciousness. Which theory we adopt will affect which types of measurement we think are valid for consciousness. This method of first identifying a theory of consciousness in humans and then trying to see where it applies in other cases - is what Birch (2020) calls the *theory heavy* approach and one which he rightly criticizes as being difficult to apply in marginal cases. Shevlin (2021) similarly describes the 'specificity problem', in attempting to specify which cognitive mechanisms in line with a particular theory of consciousness, could be applied to nonhuman cases. Even if we were to think that these problems could be solved and that identifying a correct theory of consciousness would be the best way forwards for measurement, it takes time for different theories of consciousness to duke it out amongst each other. Despite the young age of the field, it is far from clear that any consensus will emerge in the next few decades. In general, progress in philosophical theorizing is slow and often no single theory comes out as a well-supported winner (Chalmers 2015). The political and ethical issues raised by the **MPC** in identifying instances of consciousness and attributing moral status, as will be discussed in Section 5, are far too pressing to wait for such agreement.

As mentioned, it is now a common perspective to take consciousness, or sentience, to be the central factor determining the attribution of moral status (Singer 1975; Browning 2019a,2020a; Mellor and Beausoleil 2015; Duncan 2002).[5] The presence or absence of consciousness is taken to be a crucial factor in determining whether some animal should be given moral consideration, such as in recent discussions on our treatment of cephalopods (Jacquet et al. 2019; Browning 2019b; NEAVS et al. 2020) and crustaceans (Birch 2017). The very same arguments that lead us to tread carefully in our treatment of sentient and intelligent creatures, however, may eventually also be applicable to intelligent machines. As artificial intelligence grows ever more complex, one of the biggest ethical questions raised is what moral status we attribute to newly created conscious machines. Here we will not take up the debate as to whether or not machines can be conscious at all, but instead the perhaps more difficult question of how - if machines were conscious - would we detect this? (what Basl (2014) refers to as the 'epistemic challenge'). In this paper, we will focus particularly on the questions of detection of consciousness within animals and

[5] For criticism of this idea see Carruthers (2019); Dawkins (2017, 2015).

machines, arguing that due to the indicator validity and extrapolation problems, we should be very sceptical about the use of many current indicators developed using human tests in non-human animals and, in particular, in machines.

3 The Indicator Validity Problem

The first aspect of the **MPC** that we wish to discuss here is the problem of indicator validity. When measuring some target state, we can sometimes do so directly, such as when we measure the weight or length of some object, or count the number of something. However, often we do not directly measure the target and instead use indicators or proxy measures for measurement. These indicators are intended to correlate with the target, such that changes in the indicator will reflect changes in the target state. One crucial feature we require of the indicators we use is that they are valid. Validity refers to whether the indicators used are actually measuring the intended target state, as opposed to some other state (Bringmann and Eronen 2016). In most cases, when validating indicators all that is needed is to establish a correlation between the indicator and the target state. This entails measurement of both the indicator and the target under varying conditions and checking for reliable correlation (Markus and Borsboom 2013). However, in some cases, the target itself cannot be directly measured, and thus validity can not be established through these means (Browning 2020a). As mentioned previously, consciousness is one such case; its subjective nature makes it inaccessible to direct measurement and hence we have only indicators to rely on.

This then gives rise to the general problem of validity. As we cannot measure consciousness directly, we require the use of indicators or proxy measures in order to establish its presence. However, we are unable to directly validate these indicators in order to be certain that they are really tracking our intended target - consciousness. Unlike some other cases of validation, we cannot run tests with known levels of the target state - i.e. looking for indicators which are present in conscious individuals and absent in non-conscious individuals. This creates an obvious problem. Without knowing in advance which individuals are conscious and which are not, we are not equipped with any means of testing our indicators. Say we propose an indicator, and show that it is present in cases of consciousness (typically, we take humans to be a standard paradigm case). To show that the indicator is valid, we would then need to show that it is absent in cases of non-consciousness. However, without any indicators, we have no way of determining whether we have a case of non-consciousness to test with in the first place. This problem is also raised by Michel (2019), who reaches a pessimistic conclusion about our ability to ever agree on indicators for consciousness. Here, though we think there are additional reasons to be concerned about indicators in current use, we hope to provide a more positive way forward.

In cases such as these, in which we wish to validate an indicator of a hidden target, we must instead use background assumptions that link the target to the indicator (Schickore and Coko 2013). In the case of most of the currently used indicators of consciousness, this assumption appears to be that the distinction between conscious and unconscious processes in human subjects mirrors the distinction between conscious and non-conscious individuals. Here we will take non-conscious to refer to cognitive processes taking place in an individual that does not possess consciousness, and unconscious to refer to processes taking place in an individual with consciousness, but where the processes themselves are not available to consciousness. As we will describe, most current indicators are developed in humans, using differences between conscious and unconscious processing (Sandberg et al. 2010). They are then extrapolated to other cases in order to differentiate conscious from non-conscious states, under the assumption that these two distinctions are the same. As we will go on to show, this assumption lacks sufficient justification, and is likely to be incorrect.

3.1 Methods of measuring consciousness

Methods of measuring consciousness are often divided into two classes - objective and subjective (Seth et al. 2008). Objective measures look at behaviour or other observable characteristics (e.g. fMRI). An example of this is perception-based testing, in which a subject is presented with a stimulus and behavioural measures are used to determine whether it has been consciously perceived (Seth et al. 2008). Here, intentional behaviours - those behaviours that are deliberately chosen by the subject, such as selecting the next item in a sequence - are thought to demonstrate consciousness. By contrast, non-intentional characteristics of a behaviour - those effects not directly controlled by the subject, such as the speed of response; or physiological responses such as galvanic skin response - are not thought to indicate consciousness. These measures rely on the assumption that particular behaviours can only be performed in the presence of consciousness, but do not rely on self-reporting of such by subjects.

In contrast, subjective measures of consciousness are self-reports of mental states. In these tests, subjects are presented with stimuli and are asked to report on their awareness: "to ascertain whether a person knows that they know" (Seth et al. 2008, p. 317). This method can then be further fine-tuned by asking subjects to report on their confidence in their answers, or whether they believe themselves to just be guessing. In cases where subjects consider themselves to have guessed, but still perform well on tasks, they are presumed to hold unconscious knowledge. These tests rely on the assumption that self-report is a reliable guide to conscious experiencing. However, there are many reasons to be sceptical of self-reporting as an

accurate method of measurement. Firstly, an ability to report may not be directly representative of conscious experience: "without knowing what the contents of consciousness actually are, and having no other methods with which to compare introspective methods, there is no clear way of establishing when introspective errors are made, or when subjects are 'correctly' reporting their experiences" (Irvine 2012, p.634). Irvine (2012) details some of the ways in which this may occur. For instance, subjects often report being conscious of stimuli but are unable to identify or report on the details of objects or changes outside the direct field of attention. There is a risk of study subjects misreporting their experience or even confabulating their experience. Secondly, the link between recollection or reporting and conscious experience is assumed rather than established - it is possible subjects may be conscious of some stimulus in the moment but still fail to recall it or lack the ability to report it (Ginsburg and Jablonka 2019).

Indeed, many of the perceptual and cognitive processes in the human brain occur without consciousness (Dehaene et al. 2017). Tests for consciousness are thus often targeted at identifying which processes cannot be performed in this way, instead requiring conscious processing. The theory is that when certain types of processing or behaviours are found in human subjects to not be possible without conscious awareness, then these could serve as indicators of consciousness, where their presence implies the presence of consciousness. Some of the currently developed tests include masking paradigms, trace conditioning, and multisensory learning.

1) 'Masking' tests present a stimulus in such a way (usually with another stimulus immediately before and after, or with attention directed to another stimulus) as to deliberately 'hide' it from conscious perception (Kouider and Dehaene 2007). For example, a target picture (such as a picture of a face) is presented for a short period, but immediately preceded and succeeded by a picture with a repeated geometric pattern (Ginsburg and Jablonka 2019). The target picture is then 'masked', and is not consciously detected by the subject, though it can still influence subsequent behaviour and has thus been subconsciously perceived. These tests are often performed in order to determine what sorts of cognitive processing and behaviour patterns are possible with unconscious or subliminal processing, and which require consciousness.

2) Trace conditioning. Classical conditioning is a form of associative learning in which a neutral conditioned stimulus (such as the sound of a tone) is paired with an unconditioned stimulus (such as a puff of air into the eye). Conditioning is considered to have taken place where the sound of the tone then elicits the conditioned response (blinking the eye). In trace conditioning, there is a delay between presentation of the conditioned stimulus (tone) and the unconditioned stimulus (air) (Clark and Squire 1998). Trace conditioning can still result in learning only in cases where the conditioned stimulus is available to conscious processing, it:

"requires the acquisition and retention of conscious knowledge across a considerable time span" (Clark and Squire 1998, p. 79). Subjects with lesions on the hippocampus (and thus impaired memory abilities) cannot learn through trace conditioning. However, when the conditioned stimulus is masked from conscious perception (e.g. through short presentation time, or immediate 'masking' with another stimulus), then the process does not result in learning in any subjects (Knight et al. 2006; Esteves et al. 1994). It thus appears that conscious awareness of the conditioned stimulus is necessary for learning to occur. If the conditioning is successful, it is then taken to be a potential marker of consciousness, such that, when a creature is capable of learning under these conditions, it could then be presumed to be conscious.[6]

3) Multisensory learning occurs when inputs through multiple sensory streams (e.g. audio, visual, tactile) are integrated into a single cohesive sensory representation, creating a detailed representation of an object or event (Palmer and Ramsey 2012). This richer representation can then be used to guide learning. Tests using consciously and unconsciously perceived stimuli have suggested that consciousness of at least one of the primary stimuli is necessary for such learning to occur (Palmer and Ramsey 2012). Thus the ability to integrate multiple sensory stimuli and undergo multisensory learning is seen as a marker for consciousness. This is similar to the process of *Unlimited Associative Learning* (UAL) - the ability to group different sensory stimuli into a single compound percept that can be used in learning - proposed in work by Ginsburg and Jablonka as a transition marker for consciousness (Ginsburg and Jablonka 2007a,b, 2010, 2019; Bronfman et al. 2016), as will be discussed further on in this section.[7]

In general, the methods applied in these cases are to take evidence from human testing that some particular behaviour cannot be performed by humans without conscious processing. These results are then extrapolated to other cases, using the assumption that performance of this behaviour is evidence of consciousness (in humans, and also in other animals or systems). As intuitive as this assumption may be, it is not well-supported by the empirical data. As mentioned, in the first instance there are general problems with tests like these, as they may not even be accurately tracking the distinction between conscious and unconscious processing. Most often they rely on reporting by subjects of what they are or were aware of and, as discussed, these reports may not be a good guide to conscious experience. So it is entirely possible that they are not valid indicators, even for the cases in which they were developed. However, as we will argue, there is a deeper problem with these tests in that they do not take into account the distinction between unconscious and non-

[6] Though see Ginsburg and Jablonka (2019) for a criticism of trace conditioning as an indicator of minimal consciousness.

[7] See also Browning and Veit (2021) for more detailed discussion of their account.

conscious processes, and thus are highly unlikely to be valid indicators for identifying the latter.

3.2 The non-conscious/unconscious distinction

Importantly, all of the methods described above share a specific feature, giving rise to what we consider to be a particularly problematic strand of the problem of indicator validity. That is that they are all tests of the distinction between conscious and unconscious processing in human subjects. Although in many of these papers, the term 'non-conscious' is used (e.g. Kouider and Dehaene 2007), they are really describing unconscious processes. The indicators are validated on this, and though they may be valid in these cases, they are extrapolated for use in other cases only using the assumption that the distinction between conscious and unconscious processing is the same as that between conscious and non-conscious processing. This presents an important problem as there is no reason to think that what we are really tracking here is the difference between conscious and non-conscious cognition. All these tests are showing is that under the conditions tested, the unconscious processes of the brain are not capable of performing these functions. This does not show that these functions cannot be performed non-consciously in another context, such as by an organism or artificial system that lacks consciousness. The methods we have described for measuring consciousness are explicitly set up to be measuring the differences between conscious and unconscious processing in an organism that possesses consciousness. If they have been done well, we can establish these indicators as valid for distinguishing consciousness from unconsciousness in such organisms (though, as mentioned, we may have reasons to be suspicious of even these claims). However, this is insufficient to consider them valid for distinguishing between consciousness and non-consciousness. To extrapolate in this way requires a background assumption of the relationship between these two distinctions, one that we argue is not supported, for the reasons we present below.

A helpful analogy to this problem suggests itself in attempts to identify whether something is alive. In trying to determine the markers of life, we might not find it particularly helpful to examine the differences between living and dead organisms. A dead organism is still a biological organism, though one which no longer possesses life. The features of this organism, and how it differs from a living one, may not be useful in diagnosing whether some other (non-biological) entity is alive, and indeed may even be misleading. For example, dead organisms are typically undergoing decomposition[8], however we would not want to add 'decomposition' to the list of features characteristic of non-life, as it would then give us the wrong results

[8] We thank an anonymous reviewer for this example

in many cases (e.g. ruling rocks to be alive as they, like living organisms, are not undergoing decomposition). So too may tests on unconscious processes fail to give features relevant to non-consciousness.

There are a couple of reasons to doubt that the assumption of relevant similarity between the conscious/unconscious and conscious/non-conscious distinctions. One is that some processes that may require consciousness in humans could be performed non-consciously by other entities that have evolved or otherwise developed in different ways. Another is that there may be processes that are unconscious in humans (or other conscious organisms), yet cannot be performed at all by non-conscious entities because they still require the scaffolding of consciousness.

Regarding the first, we can think of the development of the capacity to perform conscious processes, both evolutionarily and ontologically. We could plausibly imagine a scenario under which the processes were once performed non-consciously, or unconsciously. As an individual develops consciousness, these formerly non-/unconscious processes move into the realm of consciousness. This could happen as a result of evolutionary change, or even developmental change over the life of an organism. Ginsburg and Jablonka (2019) give the example of an organism in which, at a young age, the food-seeking behaviour is controlled through unconscious innate processes, but as the animal grows, developing memories and learning from experiences of different foods they have encountered and their interactions with these, the behaviours become a part of a conscious process. The previous non-conscious scaffolding could be removed, such that the process now relies on conscious processing, and perhaps can *only* operate while entering conscious awareness. When tested, subjects are unable to succeed using unconscious processing. However, this does not mean that the process can't ever be done this way, only that it is not the case in these subjects. There are potentially many different mechanisms, some conscious and some not, that could be used to perform these tasks. In humans, we can see examples of such tasks in walking, or breathing, both of which can often be done unconsciously, but brought under voluntary conscious control if desired. Where some particular background conditions are present, perhaps the tasks are necessarily performed consciously, while where they are absent, the tasks are performed non-consciously. The fact that these processes are necessarily conscious in test subjects does not indicate that they are necessarily conscious in all cases.

Just because a process requires conscious processing in a conscious organism, it does not necessarily follow that it cannot also be done somehow through non-conscious means. This is particularly likely in organisms much more unlike ourselves, such as invertebrates, or machines. Given how different machines are from ourselves, it is entirely possible that they could perform tasks non-consciously that we are only

able to perform consciously (or, conversely, that they may perform tasks consciously which we do not). It is not clear how we would identify these cases, without valid indicators to use. Ginsburg and Jablonka (2019) discuss the example of 'T-robots', machines which perform sufficiently complex behaviour to pass a 'consciousness Turing test', convincing human observers that they are conscious. However, it still seems entirely possible that these robots could be non-conscious 'automata'. Although natural selection gave rise to complex conscious systems for performing complex behaviour, it is possible that artificially designed systems for the same purposes could be considerably simpler, without the requirement for consciousness. "Although a particular function may be realized by a robot that has no consciousness, in an evolved animal, it is through consciousness that this function is realized" (Ginsburg and Jablonka 2019, p. 188).

An example of such a process is the presence of a capacity for Unlimited Associative Learning (UAL), which Ginsburg and Jablonka (2019) argue could serve as a transition marker for the presence of consciousness. UAL is "open-ended learning that enables an organism to ascribe motivational value to a compound stimulus or action and use it as the basis for future learning" (Ginsburg and Jablonka 2019, p.191). It allows organisms to integrate different perceptual inputs into a single complex compound stimulus, which enhances the ability to discriminate stimuli and build associations between stimuli and actions. Ginsburg and Jablonka have not proposed this as a marker because they consider it to be a necessary or sufficient feature of consciousness, but merely an indicator to show that an organism is one that has in place the entire enabling system for consciousness. This is used as a positive marker only, with the absence not being seen as an indication of the absence of consciousness as an organism could have the enabling system but have lost UAL, or be in a stage of development where the enabling system has developed but UAL has not yet emerged. Importantly, they acknowledge that it would be possible to see UAL in a different type of system (such as a machine) without any reason to infer consciousness. Simply because a process cannot be performed unconsciously is insufficient reason to think it cannot be performed non-consciously.

The second reason for doubting the assumption of similarity between unconscious and non-conscious processes is that there could be unconscious tasks that cannot be performed non-consciously, as they require the presence of consciousness. The brain structures and processes that give rise to consciousness could be necessary scaffolding for the performance of these unconscious tasks, and thus non-conscious entities would lack the ability. Take, for instance, unconscious 'habitual' behaviours that are often taken as evidence that complex behaviours do not require consciousness. Many of us are familiar with the experience of getting in the car and driving home, only to realise when we get there that we have no memory of the actual drive – the entire process was seemingly done unconsciously. Although

driving is quite a complex behaviour, requiring integration of many perceptual inputs and responding with appropriate action patterns, it is seemingly possible to do it without conscious processing. This can be taken to undermine the claim that the ability to perform complex behaviour is good evidence of consciousness. Other examples include the complex behaviours that can be performed by somnambulists, such as reports of subjects being able to drive, shop, cook and even hold conversations while under the influence of particular sleep-inducing medications (Dolder and Nelson 2008). However, although performed unconsciously, it may be the case (as claimed by Ginsburg and Jablonka 2019), that consciousness is a necessary scaffold for these types of behaviours. That is, that unconscious habit is only possible after a behaviour has been learned and practiced, processes which require consciousness. Or the complex processing occurring during sleep is a result of conscious activity in the brain at other times. This would mean that these behaviours could only ever be performed unconsciously, not non-consciously, and thus could still serve as evidence of consciousness.

A similar example of this is the difference between limited associative learning (LAL) and unlimited associative learning (UAL) (Ginsburg and Jablonka 2019). As mentioned, UAL is taken as a marker of consciousness, because it indicates the presence of the necessary enabling system for consciousness (at least, in biological organisms). LAL, by contrast, can be performed without such an enabling system and thus by non-conscious entities. However, an organism that is capable of UAL could also experience LAL consciously: "when reflex-eliciting stimuli are processed by a brain that has an architecture that supports sentience/minimal consciousness, they become subjectively experienced because they are processed by high-level integrating units" (Ginsburg and Jablonka 2019, p.378). A process that is sometimes non-conscious can still be experienced consciously by organisms if they already have that capacity; it is not necessarily non-conscious.

The validity problem, and in particular the problem with the conflation of unconscious and non-conscious processes, will apply in any case in which indicators developed in this way are being extrapolated to try and determine whether individuals are conscious or not, including both non-human animal and machine cases. Indicators validated only on the conscious/unconscious distinction cannot be taken to be valid for detection of the conscious/nonconscious distinction, without further justification for the similarities between these. As we have argued, there are currently many reasons to think that there would be no such justification. This is particularly problematic, as it seems that almost all of the currently developed indicators use this same controversial background assumption. Thus, the convergence of evidence from these multiple sources gives us no further assurance that they are hitting the relevant target. Only through use of multiple lines of evidence coming from different sources with different and independent background assumptions (i.e. alternative models) can

we increase our confidence in the accuracy of our results (Wimsatt 2007; Veit 2020). We develop this suggestion further in the conclusion.

4 The Extrapolation Problem

Another strand of the **MPC** that will apply to measurement of non-human consciousness is the extrapolation problem. This is another form of the indicator validity problem. Whereas the problem described above related to the use of indicators validated on the conscious/unconscious distinction in cases of the conscious/non-conscious distinction; the extrapolation problem instead relates to the use of indicators validated on human subjects to nonhuman entities - animals and machines. We thus think it is worth a separate examination.

This second problem - the extrapolation problem - would arise, even if we were to solve the first. Let us imagine that we solve the validity problem to our satisfaction, that we find some indicators that we have validated in humans and we have satisfactory reasons to believe that they are measuring the right kind of distinction between conscious and non-conscious processing. If we want to then use these indicators in other (non-human) cases, we need further background assumptions to justify considering them valid for these cases. Here, we will argue that though there may be some convincing reasons to extrapolate to the case of non-human animals, these (biological) reasons will not apply to the machine case. Here we are thinking primarily of the behavioural indicators discussed earlier in the paper, though we will also briefly mention physical indicators, such as neural correlates. Indicators of consciousness must be relevant to the mechanisms through which consciousness is produced and operates, and we do not know whether these are the same in living organisms as in machines. This problem has received much attention in the recent literature (Elamrani and Yampolskiy 2019; Michel 2019; Shevlin forthcoming); here we hope to clearly outline the reasons why it is such a concern, and point to some potential solutions.

When taking indicators developed in humans and applying them to animals, we use background assumptions to justify this cross-applicability. These assumptions are that animals are relevantly similar to humans, such that the mechanisms and processes operating between the experience of consciousness and the measured indicator are likely to be of the same type. An example of this type of reasoning can be seen in the work of Berns (2018) in neuroimaging of animals: "analogous regions in dog and human brains appear to serve analogous functions. This is important, because analogous structure-function relationships provide a pathway for answering the question of what it's like to be a dog, or any other animal. I suspected that when analogous brain structures were active in an animal, they were having analogous

subjective experiences to us" (Berns 2018, p. 47). These similarities could be structural (anatomy and physiology), functional and/or historical. Unfortunately, none of these similarities apply in the machine case.

The justification for assumption of relevant background similarities between humans and animals is based in both biological analogy and shared evolutionary history. The biological mechanisms giving rise to consciousness, including perceptual apparatus, neural structures and outgoing action-driving pathways, such as hormone cascades, are analogous between many species. Additionally, as animals and humans share common ancestors, it is likely that the evolutionary pressures that gave rise to conscious experience in humans are the same as those that gave rise to this experience in other animals, long before the lineages split (for an example of such an account, see Ginsburg and Jablonka 2019). For these reasons, we feel justified in cross-applying indicators of consciousness. "Feelings, which accompany our own learning from experience, can be projected onto other animals that act like us and are anatomically similar to us; assuming they do have such feelings can provide the best explanation of the behaviours we observe, as well as predict other behaviours" (Ginsburg and Jablonka 2019, p. 196). The anatomical, physiological and evolutionary similarities provide justification for believing that there is also similarity in conscious experience. For this reason, we could feel relatively comfortable using indicators developed in humans for measurement of consciousness in non-human animals.

It is worth noting that as this confidence is based in the possession of relevant similarities, it will depend on the degree to which animals possess these similarities. As the similarities decrease, so too should our confidence in our ability to validly extrapolate our indicators. Animals more distantly related to humans, with remote common ancestors, will have a different selective history, and different neural systems. However, even in these cases, we still have some degree of biological analogy. Even where common ancestors are distant, there is still some shared evolutionary history, and thus the potential for deeper homologies at the cellular and developmental level, as well as the same selective pressures, that still give rise to relevant similarities in consciousness. Although the extrapolation problem still applies in these cases, it is weaker than in the case of machine consciousness.

The obvious problem for extrapolation to machines is that all of these lines of evidence rely on biology and thus cannot be applied to those cases. Machines have no shared history of this type with humans or other animals, and no analogous anatomical or physiological structures. Because of this, we have no reason to think that the indicators developed in humans will work for attributions of machine consciousness. The extreme disanalogies between animals and machines mean that work in one area is highly unlikely to be cross-applicable to the other. Take the capacity for UAL, as discussed in the previous section. While UAL in biological

organisms is taken as a marker of the presence of the enabling systems for consciousness, it is possible for machines to have UAL without the corresponding evolved enabling system (Ginsburg and Jablonka 2019, p. 227). It is for evolutionary reasons that the particular features of consciousness and UAL can be considered to cooccur. It is specified as a condition for consciousness only in "evolved extant animals" (Ginsburg and Jablonka 2019, p. 455), with no conclusions drawn about what it would represent in other cases.

The problem is even more pronounced if instead of these behavioural indicators, we try and take some physical or anatomical indicators, such as the neural correlates of consciousness (NCCs). These are those brain structures or pathways thought to be responsible for conscious processing. Identifying NCCs includes the use of neuroimaging techniques such as electroencephalogram (EEG), transcranial magnetic stimulation (TMS), functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), to find the neural differences between conscious and unconscious subjects and thus identify markers of consciousness. For example, binocular rivalry occurs when the two eyes are presented with different stimuli, such that the brain cannot create a cohesive visual field and instead alternates between conscious perception of each stimulus (Blake and Logothetis 2002). A similar effect can be induced aurally (Brancucci and Tommasi 2011). Neuroimaging to determine which pathways are active while each eye (or ear) gains brief conscious 'control' can help identify some of the NCCs. Where these NCCs are then present in other subjects, the implication is that they also possess consciousness. Again, extrapolation from human cases to other animals may not be well-supported (see e.g. Ginsburg and Jablonka 2019, p.98), and disagreement over such underlies recent debates about the presence of fish consciousness (Key 2016; Sneddon et al. 2014). However, even if we use the biological justifications described above to accept that these are useful markers of consciousness in biological organisms, they are too substrate- and context-specific to be considered useful across the board.

This is a version of the argument of multiple realization, i.e. whether consciousness could be instantiated through different substrates and/or mechanisms. This could occur in species that lack features some take to be necessary for consciousness in humans (e.g. the neocortex[9]), or in species that may represent an independent evolutionary origin of consciousness from our own, from a non-conscious common ancestor, e.g. cephalopods (Godfrey-Smith 2013, 2016) and arthropods (Barron and Klein 2016). It could also occur in artificial systems. While multiple realization is often raised in discussions of whether or not machines could even *be* conscious, as mentioned earlier, this is not our concern here. What is of interest is how multiple realization affects measurement of consciousness - in

[9] Though work by Merker (2007) disputes that the neocortex is necessary for consciousness.

particular how we could validate indicators developed on one type of substrate for use on others and identify which other types of entities may be conscious. As we have indicated here, this would require background assumptions that do not currently hold for the machine case. Indeed, some, such as Michel (2019) have thus argued that we could *never* empirically settle whether organisms such as fishes feel pain, but this underestimates the empirical and theoretical toolkit of biological science and in the conclusion we will suggest some possible ways forwards through this problem.

Although there may be some characteristics we determine to be necessary features of consciousness in *living* organisms, we cannot then extrapolate that these are necessary for consciousness of any type. We have an n=1 problem, where our only sample for consciousness is a related cluster of biological life. All of the cases of consciousness we know of are from a single source - evolved life on earth. Even if we take consciousness to have arisen multiple separate times within this group (e.g. Ginsburg and Jablonka 2019; Godfrey-Smith 2016), the presence of common ancestors and resulting underlying developmental and anatomical similarities that could constrain evolution for these organisms still does not give us truly independent events. Think of the evolution of animal eyes: once thought of as separate events of convergent evolution, but now understood as a result of deep homology at the level of the PAX6 gene (Shubin et al. 2009). Rather than separate evolutionary events, the emergence of consciousness may be similarly homologous and thus not represent independent data points for drawing conclusions about its necessary properties. Taking again the NCCs – even if we were to find that these were present in all conscious organisms, this would give us no reason to think they were a necessary feature of conscious processing as opposed to a contingent feature of our particular type of evolution. Although animals may be unable to perform certain processes without consciousness, this does not make consciousness necessary for these processes – it could be a contingent feature of the way consciousness has arisen within Earth-based life.

Thus, although we may go some way towards solving the **MPC** in non-human animals, due to biological similarities with humans, this will not help us in the case of extrapolation to artificial systems. At present, it does not seem that any of the indicators of consciousness developed in humans could justifiably be used for detection of machine consciousness. How, then, can we make progress on the question of machine consciousness? There are a couple of possible paths we could take in solving this problem. Firstly, we could just choose to use the indicators we have discussed above, as our best available option in the absence of other alternatives. This may be feasible in the animal case, due to analogy and shared evolutionary history, but seem much less likely to work in the machine case.

We could think that in the absence of reasons to believe otherwise, it makes most sense to accept this behavioural evidence as evidence of consciousness.

However, this only holds true if there are no defeaters undermining the link between 'conscious-type' behaviour and consciousness such that a 'common cause' would be our best explanation for the observed behavioural similarities (Michel 2019). As we have shown in this section and the one preceding, there are many reasons to resist this background assumption that similar behavioural outputs are likely to indicate the presence of similar underlying conscious processes, without the additional biological background assumptions described. At the very least, there is currently no justification for this assumption and we should remain neutral on this point. Another tactic would be to try and find other proxies we think would work in determining consciousness. If we could have at least some theoretical justification for the use of these proxies, then we could use them as a rough guide to the presence of consciousness.

Possibly the most promising solution in this vein is to move instead to functional similarities - not in terms of the function of consciousness, but in the mechanisms by which it functions. Danaher (2019) and Shevlin (forthcoming) have similarly argued that many indicators we take to work in humans are flawed when they are applied to machines. Yet, they think that behavioural or cognitive similarities are respectively the best guide towards machine consciousness. There is still a worry here, however, that such similarities can be instantiated without any similarities in the underlying mechanistic structure. We are thus doubtful that this is anything but an intermediary strategy.

One possible marker of consciousness that does not rely on biological assumptions or tests using the conscious/unconscious distinction, comes from Integrated Information Theory (IIT) (Tononi 2004). IIT looks to measure Φ, as a measure of the capacity of a system to integrate information. This measure can be used for any information processing system, as a measure of its complexity. However, it requires commitment to an integration theory of consciousness, where consciousness "corresponds to the capacity to integrate information" (Tononi 2004, p. 2), which is more of a theoretical commitment than we are willing to make at this stage. There is also the problem of determining which threshold for Φ should be set for considering when a system is conscious, and trying to establish it experimentally would run into the validation problems already described above. That is, any tests attempting to link Φ to consciousness would require prior knowledge of which systems are and are not conscious. However, with work to address these problems, this could be one of the more promising measures of consciousness for machines.

There may be some other functional descriptions of consciousness that can lay out the requirements for consciousness – e.g. 're-entrant signalling' (linking connections between different parts of the brain and body that signal back and forth and create the integration that is often considered necessary for consciousness (Ginsburg and Jablonka 2019, p. 122). However, identifying necessary features of

consciousness will not always help in indicating the presence of consciousness. While their lack demonstrates the absence of consciousness[10], unless they are also sufficient for consciousness, their presence will not demonstrate the presence of consciousness.

Other more functional approaches to identification of consciousness may also help overcome many of these problems. For example, Ginsburg and Jablonka (2019), describe the functional composition of a conscious processing system, and the necessary structural units (e.g. hierarchical levels and general-purpose integration units). These are not specified in terms of any biological substrates, and so may be useful starting points for detection of consciousness – either through claiming that any system containing these units would be conscious, or that any system without them would not be (or both). This is still not definitive, as the presence of these processing units could be a contingent feature of biological consciousness (with consciousness possible without them) or could be a necessary but not sufficient feature, with their presence being compatible with non-consciousness. However, in the absence of better measures, we may just treat one of these accounts as the best present proxy-measure available.

5 The Moral Problem

All of this leaves us with a serious intermediate problem. It seems we are unable to use our current indicators for detection and measurement of nonhuman consciousness, especially in machines. However, if it is possible that machines could be conscious, then we will still have to make decisions regarding their treatment. As discussed in Section 2, where we take consciousness to ground moral status, then how to act in these cases of uncertainty is a pressing moral problem. In the previous section, we looked at some suggestions for using indicators developed using the functional similarities in consciousness, rather than similarities in structure or development. Though we were sceptical that these could be considered highly reliable, we acknowledged that they may be the best methods currently available to us. However, when thinking of the moral problem, what we take to be the bigger problem with most of these accounts is that they may not give us information about the right types of consciousness.

There are multiple proposed types of consciousness, such as phenomenal and access consciousness (Block 1995), self-awareness (Carruthers 2003) and cognitive and affective consciousness (Panksepp 2005). However, only one of these concepts is typically considered relevant to moral status - affective consciousness (the capacity

[10] At least insofar as we think the identified features really are *necessary* for consciousness and not simply those found to be required for consciousness in humans.

to experience positively and negatively valenced mental states, such as pleasure and suffering). Although in evolved organisms, it is considered that perceptual and affective experiences are integrated into a single conscious experience (due to the requirement for categorising and assessing the value of perceptual input) (Ginsburg and Jablonka 2019, p. 380), this does not mean that they cannot come apart, particularly within artificial systems.

It is possible that machines could possess other types of consciousness without possessing affective consciousness. They could have awareness of stimuli, and conscious cognitive processing, without any accompanying 'feeling' that any experienced states or stimuli were good or bad. However, in this case, we would not consider that there was a moral concern about the experiences of this machine. Such a machine would be unable to experience pleasure or suffering. "If we create a consciousness with only the capacity for experiencing colors but with no attending emotional or other cognitive response, we need not worry about wronging said consciousness" (Basl 2014, p. 84). It is the positive or negative valence that adds moral weight to conscious experience, and so for the purposes of ethical consideration, it is this capacity we are concerned with.[11] Thus, measures of consciousness such as those above may not be necessarily measures of affective consciousness. It is entirely possible, for instance, for a system to have a high Φ score, and perhaps have some sort of perceptual or access consciousness, without also having any affective experience of the type we are concerned with. We should, therefore, also make sure that we have a measure of the right type of consciousness so that we are able to accurately identify our targets of moral concern.

One functional account of affective consciousness postulates that it plays a role in learning and motivation – in categorising stimuli and motivating action to attain or avoid these (Cabanac 1992; Fraser and Duncan 1998; Ginsburg and Jablonka 2019). This might lead to the thought that motivation testing could be a sign of consciousness. If a system shows motivation to work in order to avoid some state of affairs and to achieve others, perhaps this could reliably signal that it feels positively about one and negatively about another. However, this proposal does not seem like a promising one. In the first instance, the link between affect and motivation might be only an evolved one, such that working to avoid things that feel bad might not be a necessary feature of machine consciousness, particularly where there is perhaps insufficient programming to allow for this. Further, there are plenty of non-conscious organisms that will avoid noxious stimuli (e.g. bacteria moving away from acid or plants growing towards the sun). The presence of consciousness is neither necessary

[11] This does not entail that any human that lacks the ability to feel pain (i.e. suffers from congenital analgesia) is no longer a subject under moral consideration as they would still experience other negatively valenced states.

nor sufficient for this type of behaviour. A more complex version of this may be the ability to make motivational tradeoffs (Ginsburg and Jablonka 2019; Spruijt et al. 2001), however, although this may not be possible for evolved organisms without consciousness, it seems that we could imagine non-conscious value-weighted programming that could achieve the same ends in machines, similar to the discussion on UAL above.

Given our lack of any valid measures of affective consciousness that could be applied in the machine case, another option is to simply wait until our consciousness science improves, until we know more about the function and mechanisms of consciousness, so that we have more confidence in our indicators. However, as this may mean we fail to recognise particular groups of conscious animals, or the advent of conscious machines, we risk doing harm to these individuals as we wait. Agar (2019) similarly argues, that even if our current theories of mind suggest that machines are unconsciouss, "treating them as if they are mindless risks wronging them" (p.1).

This may then call for the application of a precautionary principle in the face of our limited evidence of consciousness. Birch (2017) argued for such a position in the case of non-human animals whose consciousness has not been established as of yet. Precautionary principles state that under conditions of uncertainty, we should err on the side of avoiding harm. The idea in this case being that, the potential harm we could do by ignoring evidence of consciousness if it is present is far worse than the costs of mistakenly attributing consciousness where it is absent. If we apply Birch's precautionary principle and find strong evidence for the consciousness in a particular species, we would be well advised to treat the entire order as conscious, rather than demand empirical evidence for each species before they are granted protection.[12] Such an argument grounded in homology cannot be made for intelligent machines. Nevertheless, in cases where we are unsure, but have some reason to think we are dealing with a conscious machine, we may be better placed to extend ethical concern rather than to withhold it. This may be taken as too demanding, but there are others, such as Agar (2019), that have made an even stronger point, arguing that even if we have a high degree of confidence in the lack of minds in machines, we should be reluctant to treat them as such. Far from extreme or scientifically dubious, our argument is thus merely placed in a long tradition of analogous inductive-risk problems found in science.

[12] Here, anectotal evidence has a potentially useful role to play (Browning 2017).

6 Conclusion

In this paper, we have given many reasons to be sceptical about the measurement of non-human consciousness, particularly in the case of potential machine consciousness. However, as indicated, this does not mean we should do nothing. Instead, the best we may be able to do is to use the imperfect measures we already have, while staying alert as to the limitations of these methods, as we have described in the preceding sections. This paper could be read as quite pessimistic; in particular the indicator validity problem may suggest that we could not validate any indicator measures of consciousness due to our inability to directly access or measure it. Indeed, this is the conclusion of Michel (2019) who thinks that as current neuroscientific evidence is underdeterminate and cannot rule definitively in favour of or against consciousness in fish, we should remain agnostic.

However, we don't see the problems we have presented as fatal for the measurement of consciousness (see also Veit and Huebner 2020). Instead, we advocate a more pluralistic method, using something like robustness analysis, where a result that remains invariant under different experiments, models, and background assumptions can be considered more robustly supported (Wimsatt 2007; Veit 2020). In this case, we suggest that with multiple lines of evidence, tested against one another, and using independent background assumptions, we can infer that the best explanation for observed differences is differences in consciousness (Browning 2020a,b; Veit & Browning 2020a, forthcoming). In particular, all should not rely on the background assumption of the similarity across the unconscious/non-conscious distinction, or on biological similarity. While presence of one type of evidence for consciousness may be considered insufficient, when more strands are added we can gain confidence. For example, we might have the presence of trace conditioning behaviour, avoidance of noxious stimuli and the presence of appropriately complex neural processing units (or circuitry). Though each of these types of evidence alone may be poorly supported, when taken together they are stronger and in this case we would infer the presence of consciousness as the best explanation of our results.

Even where we may never have complete confidence in the validity of our measures, we can still identify which we consider to be more or less valid, based on the degree to which they have been tested and how well independent measures may correlate with one another (Browning 2020a). This is in line with the 'facilitation hypothesis' advocated by Birch (2020). Here, we take an assumption that consciousness facilitates a "cluster of cognitive abilities" and can thus look for the presence or absence of a number of abilities within this cluster to strengthen or weaken the case for the presence of consciousness in some entity under investigation.

Even if we are never able to completely solve the **MPC**, we hope that identification of and discussion of these problems, as we have done here, can assist

in improving the ways in which we measure consciousness and the inferences we make from our tests. Being able to identify the limitations of our methods, and the background assumptions needed to justify their use, allows us to then work to overcome them. In particular, the ability to identify tests with independent background assumptions and to use them together will strengthen our confidence in the results of consciousness science. This paper is thus not intended to impede research into the measurement of consciousness. Rather, we see it as a call to analyse and improve our methods accordingly - especially in the burgeoning field of artificial intelligence and the evolution of sentience - with the hope that one day we may be able to solve the ***measurement problem of consciousness***.

References

Agar, N. (2019). How to treat machines that might have minds. *Philosophy & Technology*, *33*, 269-282.

Barron, A.B. and Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences, 113*(18), 4900–4908.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology, 27*(1), 79–96.

Berns, G. (2018). *What It's Like to Be a Dog: And Other Adventures in Animal Neuroscience*. Oneworld Publications.

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience, 16*(1).

Birch, J. (2020). The search for invertebrate consciousness. *Noûs*, 1-21.

Blake, R. and Logothetis, N.K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*(1), 13–21.

Block, N. (1995). Some concepts of consciousness. *Sciences, 18*(2), 1–28.

Block, N., Carmel, D., Fleming, S.M., Kentridge, R.W., Koch, C., Lamme, V.A., Lau, H., and Rosenthal, D. (2014). Consciousness science: real progress and lingering misconceptions. *Trends in Cognitive Sciences, 18*(11), 556–557.

Brancucci, A. and Tommasi, L. (2011). "Binaural rivalry": Dichotic listening as a tool for the investigation of the neural correlate of consciousness. *Brain and Cognition, 76*(2), 218–224.

Bringmann, L.F. and Eronen, M.I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology, 26*(1), 27–43.

Bronfman, Z.Z., Ginsburg, S. and Jablonka. E. (2016). The transition to minimal consciousness through the evolution of associative learning. *Frontiers in Psychology 7*, 1954

Browning, H. (2017). Anecdotes can be evidence too. *Animal Sentience, 2*(16), 13.

Browning, H. (2019a). The natural behavior debate: Two conceptions of animal welfare. *Journal of Applied Animal Welfare Science*, 1–13.

Browning, H. (2019b). What should we do about sheep? The role of intelligence in welfare considerations. *Animal Sentience, 4*(25), 23.

Browning, H. (2020a). *If I Could Talk to the Animals: Measuring Subjective Animal Welfare.* Ph.D. thesis, Australian National University. https://doi.org/10.25911/5f1572fb1b5be

Browning, H. (2020b). Assessing Measures of Animal Welfare. *Preprint.* http://philsci-archive.pitt.edu/17144/

Browning, B. and Veit, W. (2021). Evolutionary biology meets consciousness: Essay review of Simona Ginsburg and Eva Jablonka's The Evolution of the Sensitive Soul. *Biology and Philosophy.*

Cabanac, M. (1992). Pleasure: The common currency. *Journal of Theoretical Biology, 155*(2), 173–200.

Carruthers, P. (2003). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

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

Chalmers, D.J. (2015). Why isn't there more progress in philosophy? *Philosophy, 90*(1), 3–31.

Clark, R.E. and Squire, L.E. (1998). Classical conditioning and brain systems: The role of awareness. *Science, 280*(5360), 77–81.

Danaher, J. (2019). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, 1–27.

Dawkins, M.S. (2015). Animal welfare and the paradox of animal consciousness. *Advances in the Study of Behavior*, *47*, 5–38.

Dawkins, M.S. (2017). Animal welfare with and without consciousness. *Journal of Zoology, 301*(1), 1–10.

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492.

Dolder, C.R. and Nelson, M.H. (2008). Hypnosedative-induced complex behaviours. *CNS Drugs, 22*(12), 1021–1036.

Duncan, I.J. (2002). Poultry welfare: Science or subjectivity? *British Poultry Science, 43*(5), 643–652.

Elamrani, A. and Yampolskiy, R.V. (2019). Reviewing tests for machine consciousness. *Journal of Consciousness Studies, 26*(5-6), 35–64.

Esteves, F., Parra, C., Dimberg, U., and Ohman, A. (1994). Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology, 31*(4), 375–385.

Fraser, D. and Duncan, I.J. (1998). 'Pleasures', 'pains' and animal welfare: Toward a natural history of affect. *Animal Welfare, 7*(4), 383–396.

Ginsburg, S. and Jablonka, E. (2007a). The transition to experiencing: I. Limited learning and limited experiencing. *Biological Theory, 2*(3), 218– 230.

Ginsburg, S. and Jablonka, E. (2007b). The transition to experiencing: II. The evolution of associative learning based on feelings. *Biological Theory, 2*(3), 231–243.

Ginsburg, S. and Jablonka, E. (2010). The evolution of associative learning: A factor in the Cambrian explosion. *Journal of Theoretical Biology, 266*(1), 11–20.

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

Godfrey-Smith, P. (2013). Cephalopods and the evolution of the mind. *Pacific Conservation Biology, 19*(1), 4.

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux.

Irvine, E. (2012). Old problems with new measures in the science of consciousness. *The British Journal for the Philosophy of Science, 63*(3), 627–648.

Jacquet, J., Franks, B., Godfrey-Smith, P. and Sánchez-Suárez, W. (2019). The case against octopus farming. *Issues in Science and Technology, 35*(2), 37–44.

Key, B. (2016). Why fish do not feel pain. *Animal Sentience, 1*(3), 1–34.

Knight, D.C., Nguyen, H.T. and Bandettini, P.A. (2006). The role of awareness in delay and trace fear conditioning in humans. *Cognitive, Affective, & Behavioral Neuroscience, 6*(2), 157–162.

Kouider, S. and Dehaene, S. (2007). Levels of processing during nonconscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 857– 875.

Markus, K.A. and Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning* (1st ed.). New York: Routledge.

Mellor, D.J. and Beausoleil, N.J. (2015). Extending the 'Five Domains' model for animal welfare assessment to incorporate positive welfare states. *Animal Welfare, 24*(3), 241–253.

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

Michel, M. (2017). Methodological artefacts in consciousness science. *Journal of Consciousness Studies, 24*(11-12), 94–117.

Michel, M. (2019). Fish and microchips: on fish pain and multiple realization. *Philosophical Studies, 176*(9), 2411–2428.

Michel, M., Fleming, S.M., Lau, H., Lee, A.L., Martinez-Conde, S., Passingham, R.E., Peters, M.A., Rahnev, D., Sergent, C. and Liu, K. (2018). An informal internet survey on the current state of consciousness science. *Frontiers in Psychology, 9*, 2134.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical review, 83*(4), 435–450.

New England Anti-Vivisection Society (NEAVS) et al. (2020). Petition to Include Cephalopods as "Animals" Deserving of Humane Treatment under the Public Health Service Policy on Humane Care and Use of Laboratory Animals. *Harvard Law School Animal Law & Policy Clinic*, 1–30. https://doi.org/10.13140/RG.2.2.27522.30401

Palmer, T.D. and Ramsey, A.K. (2012). The function of consciousness in multisensory integration. *Cognition, 125*(3), 353–364.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition, 14*(1), 30–80.

Papineau, D. (2003). Could there be a science of consciousness? *Philosophical Issues, 13*, 205–220.

Rosenthal, D. (2019). Consciousness and confidence. *Neuropsychologia, 128*, 255–265.

Sandberg, K., Timmermans, B., Overgaard, M. and Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Consciousness and Cognition, 19*(4), 1069–1078.

Schickore, J. and Coko, K. (2013). Using multiple means of determination. *International Studies in the Philosophy of Science, 27*(3), 295–313.

Seth, A.K., Dienes Z., Cleeremans, A., Overgaard, M. and Pessoa L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences, 12*(8), 314–321.

Shevlin, H. (forthcoming). How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics*.

Shevlin, H. (2021). Non-human consciousness and the specificity problem: a modest theoretical proposal. *Mind and Language*, 1-18

Shubin, N., Tabin, C. and Carroll, S. (2009). Deep homology and the origins of evolutionary novelty. *Nature, 457*(7231), 818–823.

Singer, P. (1975). *Animal liberation: A new ethics for the treatment of animals*. London: Jonathan Cape.

Skinner, B. (1953). *Science and Human Behavior*. Macmillan.

Sneddon, L. U., Elwood, R.W., Adamo, S.A. and Leach, M.C. (2014). Defining and assessing animal pain. *Animal Behaviour, 97*, 201–212.

Spruijt, B.M., van den Bos, R. and Pijlman, F.T. (2001). A concept of welfare based on reward evaluating mechanisms in the brain: anticipatory behaviour as an indicator for the state of reward systems. *Applied Animal Behaviour Science, 72*(2), 145–171.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42.

Van Gulick, R. (2018). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition ed.).

Veit, W. (2020). Model Pluralism. *Philosophy of the Social Sciences, 50*(2), 91–114.

Veit, W. & Browning, H. (2020). Perspectival pluralism for animal welfare. *European Journal for Philosophy of Science,* 11(9). https://doi.org/10.1007/s13194-020-00322-9

Veit, W. & Huebner, B. (2020). Drawing the boundaries of animal sentience. *Animal Sentience* 29(13).

Veit, W. and Browning, H. (forthcoming). Phenomenology Applied to Animal Health and Suffering. In S. Ferrarello (Ed.), *Phenomenology of Bioethics: Technoethics and Lived Experience*. Springer. http://dx.doi.org/10.13140/RG.2.2.31185.76645

Watson, J.B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158–177.

Wimsatt, W. C. (2007). *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.