Do current evidential standards in the science of consciousness help or hinder the discovery of signs of consciousness?

Uwe Peters
Utrecht University
[This is a draft. Comments very welcome!]

Abstract

One recent approach to determining whether comatose patients, non-human animals, or brain organoids are conscious is to examine whether they display features that scientific studies have found to be correlated with and indicative of consciousness. However, it is unclear to what extent scientific studies that search for such signs of consciousness rely on evidential standards that facilitate the detection of these features. Here, I argue that when it comes to standards of statistical significance, many of the studies at issue rest on a value judgment according to which false positive research conclusions are much more problematic than false negative ones. This value judgment contradicts a common normative intuition that many consciousness researchers have and may impede the discovery of signs of consciousness. Moreover, recent efforts to reduce replication failures of scientific studies by lowering the threshold for statistical significance may further increase the risk of consciousness researchers to miss evidence of consciousness in organic or artificial systems. I argue that these limitations provide reasons to shift from the conventional statistical significance thresholds in experimental consciousness research to Bayes Factors.

Keywords: consciousness; false negatives; false positives; evidential standards; statistical significance

1. Introduction

Consciousness, here construed as an organism's subjective experience ('phenomenal consciousness') of the world or its own body (Block, 1995), is commonly taken to be the main source of much of what is valuable in the world (Cleeremans & Tallon-Baudry, 2022). For instance, it is often because we experience some activity as pleasurable that we treat it as valuable, and organisms that are conscious are commonly thought to deserve legal protection from harm precisely because they are conscious (Levy, 2014). Correspondingly, it has been noted that the line between experiencing beings and others is "arguably the most important theoretical line to be drawn in the whole of reality" (Strawson, 1994, p. 154).

Yet, while we have everyday measures for determining whether a being is conscious, for example, verbal reports of experience or voluntary action, these measures cannot readily be applied to challenging cases such as preverbal infants, vegetative patients, non-human animals, or brain organoids (Farisco et al., 2022). These beings or systems typically cannot report their potential experiences or may not display voluntary actions associated with consciousness. However, determining whether consciousness is present is in these cases especially important because significant harm may result if these organisms or systems are treated as unconscious when they are in fact conscious (e.g., vegetative patients that are mistakenly assumed to be permanently unconscious may have their life support withdrawn) (Rosanova et al., 2012). Call the problem of determining whether a particular (organic or artificial) system in these challenging cases is conscious the *detection problem*.

One promising recent approach that a number of philosophers of science and consciousness researchers have adopted to address this problem is to avoid looking for a single key indicator or symptom of consciousness. Instead, the idea is to catalogue the various dispositions or capacities associated with consciousness (including certain kinds of learning, neural activation patterns, agency, etc.), determine whether they reliably cluster together, and, if so, treat them or the property (or properties) that produce this clustering as a measure or marker of consciousness in challenging cases (Shea & Bayne, 2010; Bayne & Shea, 2021, Birch, 2022; Dung & Newen, 2023). This 'cluster' approach[1] can yield stronger support for inferences about consciousness than reliance on a single sign of consciousness does because by pointing to an underlying potential natural kind, the discovery of a cluster gives each individual consciousness indicator, which may be only weakly apparent in a particular test, greater evidential weight than it would have in isolation (Bayne & Shea, 2021). Correspondingly, this approach to detecting consciousness does not just rely on findings from a single experimental paradigm but draws on converging evidence aggregated from many different behavioural, psychological, and neurobiological studies. I'm sympathetic to the cluster approach. The focus here will be on the methodology of the scientific studies of consciousness that this approach to the detection problem relies on.

There is a key challenge that these studies face when it comes to determining whether a hypothesis about consciousness is true. Specifically, due to its probabilistic nature, scientific evidence never suffices to completely prove or disprove a given hypothesis (Douglas, 2009). There always remains an "inductive risk", i.e., a risk of accepting "false positives" (concluding that there is an effect when there is not) or "false negatives" (concluding that there is no effect when there is one) (Elliot & Richard, 2017). Correspondingly, when scientists make ascriptions of consciousness, they may mistakenly judge an unconscious organism or process to be conscious (false positive) or a conscious organism or process to be unconscious (false negative) with potentially harmful consequences (e.g., in clinical domains; Peterson et al., 2015; Birch, 2023). When formulating and applying criteria for the ascription of consciousness, researchers therefore need to consider the relative costs of these two forms of error and balance them off against each other, which is "not a purely scientific task but requires attention to complex and contested ethical questions" (Shea & Bayne, 2010, p. 463).

The way researchers balance the risk of false positives against the risk of false negatives commonly manifests in the threshold for statistical significance (i.e., a p-value < 0.05) that they set for their results and that thus specifies what counts as sufficient evidence for accepting a claim (Magnus, 2022). Different consciousness researchers deal differently with this threshold, leading in some cases to disagreements (Birch, 2023). For instance, when Cruse et al. (2011) reported that a machine learning model using EEG data from healthy individuals and vegetative patients found that several outwardly unresponsive, vegetative patients were reliably responding to commands to imagine

[1] The approach is diverse. For instance, Bayne and Shea's (2021) "natural kind" version of this approach rests only on assuming that consciousness is a natural (homeostatic property) kind, whereas Birch's (2022) "facilitation hypothesis" version, which assumes that consciousness has behavior/cognition-facilitating function, is more committal.

acting in certain ways, Goldfine et al. (2013) replied that Cruse et al. failed to adjust their significance threshold to multiple testing, resulting in false positives in consciousness ascriptions to the patients. In response, Cruse et al. (2013) rejected the proposal that their significance threshold was too permissive and held that conservative corrections to it would unacceptably increase false negatives (i.e., missed responsiveness). There are more recent instances of such disagreements on how to balance false positives against false negatives in this domain (Claassen et al., 2019; Birch, 2023).

However, none of these contributions has yet examined whether the current statistical thresholds that studies on consciousness conventionally accept already rest on a particular value judgment concerning the relative risks of false positives and false negatives. This is problematic because it might be that these evidential standards are based on an implicit value judgment that makes the discovery of signs, symptoms, or measures of consciousness more difficult than it could be. Since finding ways of detecting consciousness in challenging cases is urgent, it is vital to investigate this question. The goal here is to do so.

I will argue that many scientific studies that investigate whether a given behavioral, cognitive, or neural feature correlates with, facilitates, or depends on consciousness rely on evidential standards of statistical significance that rest on the value judgment that false positives are much more problematic than false negatives. This judgment contradicts a common ethical intuition that many consciousness researchers have. It may also hinder the discovery of signs of consciousness by contributing to an oversight of cases in which behavior, cognition, or neural features are associated with consciousness. Moreover, recent efforts to reduce replication failures in science by making benchmarks for statistical significance more stringent (Benjamin et al., 2018) can increase the chances of such an oversight. I end by discussing mitigation strategies and suggest that these ethical and epistemic problems provide grounds for replacing conventional significance benchmarks and classical hypothesis testing in consciousness science with Bayesian testing and Bayes Factors.

I begin by arguing that many researchers working on the detection problem have recently advocated the view that false negatives in consciousness ascriptions are prima facie more problematic than false positive ascriptions. I will then contend that the evidential benchmarks of many studies on consciousness fail to capture this normative intuition and in fact rest on the opposite value judgment.

2. A common intuition about consciousness ascriptions

Many researchers working on the detection of consciousness in challenging cases hold that false negatives in consciousness ascriptions are especially important to avoid (Bradshaw, 1998; Fins & Bernat, 2018; Niikawa et al., 2022). The basic rationale is that when a being or system *A* is mistakenly viewed as not conscious even though it is conscious, this would cause harm to *A* rather than benefit it, depicting *A* as lacking a feature that it in fact has. In contrast, when *A* is mistakenly viewed as conscious even though it is not conscious, this would not cause harm to *A*. Overall, then, mistakenly viewing *A* as not conscious, i.e., a false negative, would bring about more harm than benefit to *A* (Niikawa et al., 2022).

False *positives* in consciousness ascriptions, too, may also cause harm, for instance, to other individuals. When laboratory rats are ascribed consciousness despite lacking it and so are no longer used for cancer research, this may harm cancer patients. Similarly, misattributions of responsiveness to vegetative patients may lead to patients being kept alive for longer when their prospects of recovery to a level that they would themselves want are grim (Birch, 2023). Since false positives in consciousness research can therefore create potentially significant ethical costs too, I will remain agnostic here on whether either false negatives or false positives in this domain are overall more problematic.

My point is simply that the ethical intuition that false negatives in consciousness ascriptions are prima facie more undesirable than false positives is common among consciousness researchers. For instance, in the literature on animal consciousness, many researchers advocate a "precautionary principle", which captures a "better safe than sorry" position: When considering whether an animal feels pain and when encountering mixed evidence that it may do so, we should give the animal the benefit of the doubt such that in the absence of strong counterevidence, we should treat it as sentient (Bradshaw, 1998; Jones, 2016; Brown, 2016; Seth, 2016; Birch, 2017). Relatedly, in research on brain organoids, Koplin and Savulescu (2019) propose that "brain organoids should be screened for advanced [potentially consciousness indicating] cognitive capacities they could plausibly develop. In general, such assessments should err on the side of overestimating rather than underestimating cognitive capacities" (p. 765). Building on this notion, Niikawa et al. (2022), too, argue that "if we are not certain whether brain organoids have consciousness – and where treating [these organoids] as not having consciousness may cause harm to them – we should proceed as if they do have consciousness" (p. 1). Hence, many consciousness researchers view false negatives in consciousness ascriptions as worse than false positives. That is, they would favor overascriptions of consciousness to underascriptions.

To what extent do the evidential standards for accepting or rejecting a claim in scientific studies of consciousness align with this common ethical intuition? Advocates of the views just outlined have so far not explicitly considered this question. But some have suggested that the tendency to view false negatives in consciousness ascriptions as more problematic than false positives should at best only influence policymaking (e.g., about animals or brain organoids), not the evidential standards of science. For instance, Birch (2017), who defends one version of the precautionary principle in the context of animal consciousness, holds that while a "low evidential bar […] should be applied when making a precautionary attribution of sentience on the basis of a single credible indicator and when extrapolating across a whole order from a single species", there "should not be any lowering of standards with regard to the methodology of experiments, or with regard to the analysis of experimental data" (p. 8). Birch suggests that we should retain the current scientific standards because otherwise critics of the precautionary principle may maintain that applying this principle means researchers' subjective values skew their studies toward overascriptions of consciousness to animals. The thought is that lowering the scientific standards to better align them with the intuition that false negatives are especially problematic in consciousness ascriptions should be avoided because it risks undermining people's trust in the objectivity of studies on consciousness.

While I share this concern, it can give the impression that the current evidential standards in consciousness studies are ethically neutral, or that they treat false positives and false negatives in consciousness ascriptions as equally problematic (to prevent biasing inquiries either way). But is this the case? Finding an answer is important. For suppose these standards are not ethically neutral, do not treat both errors equally, or, for instance, treat false positives as worse than false negatives without it having been shown that the overall costs of the former type of error are higher than those of the latter. If that were so, then opponents of value-laden science might equally well cite this imbalance, too, to support their potential distrust in the objectivity of consciousness studies because a preference for either kind of error could in this case unduly interfere with impartiality.

In the remainder, I argue that the current evidential standards in many studies of consciousness are in fact not neutral but based on the opposite normative intuition than the common view that false negatives in consciousness ascriptions are more undesirable than false positives. To provide the background for the argument, I begin with a brief introduction to two key evidential standards for statistical inferences in the sciences in general.

3. Two conventional benchmarks for false positives and false negatives

To illustrate what evidential standards many scientists, in general, use to deal with the risk of false positives, also known as 'Type I errors', or false negatives, also known as 'Type II errors' (Elliot & Richard, 2017), consider an example. Suppose a group of oncologists wants to examine the carcinogenic effects of dioxin on liver tissue. They therefore expose one of two groups of rats to dioxin, later take samples from the rats' liver tissue, and then compare the rates of cancer in their experimental and control groups of rats. Suppose the first group's rate is higher. To check whether this difference is only a fluke and so a false positive, it remains "methodological orthodoxy" in science to statistically test the "null hypothesis" ('H0'), i.e., the thesis that there is no group difference (Mudge et al., 2012). This happens by calculating the results' *p*-value, which indicates the probability of the observed group difference or a larger one if the H0 is true and the test were frequently repeated (hence 'frequentist statistic') (Hoijtink et al., 2019). Importantly, since the *p*-value is between 0 and 1, scientists need to define in advance what value counts as sufficient to hold that the finding is not due to chance but statistically significant (Di Leo & Sardanelli, 2020). In defining this threshold, they set the specific Type I and Type II error probabilities that they are willing to tolerate.

The Type I error chance is called the 'alpha level' or 'α'. Across the sciences, with only rare exceptions (Ioannidis, 2019), it is "canonical" to set α to 0.05 or less, meaning that the chances will be 5% (1 in 20) or less that an observed finding is a false positive. This threshold is arbitrary in that a higher or lower value could equally well be adopted (Nuzzo, 2014), and statistics textbooks commonly note that there is "no correct alpha level" (Katz, 2006, p. 132). Nonetheless, 0.05 is the currently still pervasive benchmark (Wasserstein et al., 2019).

In addition to setting a low α to avoid Type I errors, researchers also need to make their study sensitive enough to detect a real effect if there is one, i.e., they also need to avoid Type II errors (Fiedler et al., 2012). The measure of how small an effect an experiment can detect is called "power" and specified as 1 – beta or 'β', where β is the Type II error probability. Scientists, including clinical researchers,[2]commonly set their β level, i.e., their Type II error threshold, to 0.20, which gives them an 80% chance (power) to detect an effect if there is one (Gupta et al., 2015). But just as the α of 0.05, this β, too, is an arbitrary, "conventional" benchmark (Cohen, 1988, p. 56). While researchers may aim for higher power and a lower β, study power is related to sample size such that, with a fixed α, higher power requires larger samples. This partly explains why the currently most common β is 0.20 (80% power), because large samples are often harder to obtain due to, for instance, researchers' resource limitations (Lakens et al., 2018).

4. Do studies of consciousness also rely on the conventional evidential standards?

Studies of consciousness are diverse and may explore, for instance, whether a being is globally conscious (e.g., wakeful awareness or dreaming versus comatose state), how it is conscious locally once global consciousness is present (e.g., perceptual experiences, or bodily sensations), or what it is (or is not) conscious of (conscious versus unconscious content or stimulus processing) (Bayne et al., 2016). Researchers who work on the detection problem often cite studies of local states and processes to identify dispositions and capacities associated with consciousness and obtain indicators of the presence of global consciousness (Bayne & Shea, 2021; Birch, 2022). I will here do likewise.

Focusing on these studies, while there are other methods for studying consciousness (e.g., first-person data, single-subject studies, etc.) than through group comparisons and H0 significance testing (Chalmers, 2013), the two evidential benchmarks outlined in the preceding section are also applied in many experiments that investigate consciousness. Frequently, in studies on the effects of consciousness, two or more groups of participants are presented with a stimulus either consciously or unconsciously and then their responses in a stimulus-related task are measured to determine whether the conscious processing of the stimulus played a role in facilitating the response (Dehaene & Changeux, 2011). Potential group differences are tested against the H0 that the conscious processing played no role (Vadillo et al., 2016).

Consider, for instance, studies that researchers have cited to illustrate how the scientific identification of consciousness effects could help address the detection problem. Shea and Bayne (2010) and Birch (2022) mention experiments by Perruchet (1985) and Clark and Squire (1998, 1999). In these experiments, participants could form an association between a tone and a puff of air to the eye (so that the tone caused an eye blink) when a puff of air to the eye was administered during the tone occurrence ('delay conditioning'). However, for 'trace conditioning' – when the air puff occurs shortly after the tone has stopped – consciousness of the contingency between tone and air puff was needed. Shea and Bayne, and Birch therefore propose that trace conditioning may be an indicator of consciousness. The key point here is that the mentioned experiments used H0 testing with $\alpha = 0.05$.

Studies on implicit learning also routinely depend on null results measured against this α level (Vadillo et al., 2016). In a standard experiment on implicit processing involving subliminal perception, participants' performance on a particular task is usually found

[2] Unfortunately, many scientists do not consider the power of their studies; see Szucs and Ioannidis (2020).

to be above chance but this performance is often not accompanied by consciousness of the environmental cues that caused the behavior (Dehaene et al., 1998). Crucially, to assess the absence of stimulus awareness, in these studies, researchers examine whether participants fail to detect the relevant stimuli (i.e., perform at chance in a recognition task), and "learning is assumed to be unconscious if a statistical comparison yields a null result" ($p > 0.05$) in such checks (Vadillo et al., 2016, p. 89). While it has been noted that this inference conflates the absence of evidence of consciousness with evidence of absence of consciousness (Dienes, 2015), what matters here is just that researchers who investigate implicit learning, too, typically set their α to 0.05 and base 'yes–no' judgments on the presence or absence of conscious processing on this benchmark.

In fact, some journals dedicated to publishing work on consciousness and its links to behavior and cognition require that studies using classical hypothesis testing adopt this conventional α. For instance, the American Psychological Association (APA) journal *Psychology of Consciousness* states that for "submissions that propose frequentist testing of a primary hypothesis (H1) against a null hypothesis (H0), the APA requires that the alpha level – the likelihood of accepting H1 when H0 is true – be .05 or less".[3] Hence, consciousness researchers using H0 testing cannot easily adopt a laxer α, because a *p*-value < 0.05 for statistical significance is currently "a requirement for publishing in a top journal" across the sciences (Vidgen & Yasseri, 2016, p. 1).

Finally, turning from α to β, there is ground to believe that many studies of consciousness also set their β to the conventional 0.20 or higher, corresponding to 80% or lower power. In fact, there is evidence that neuroscientific studies, including experiments on consciousness, commonly only have much lower power because their samples are notoriously too small (Button et al., 2013). While researchers' resource constraints are one hindrance to recruiting large samples, in consciousness studies using invasive tests on animals (Mazor et al., 2023), there is yet another problem: Larger samples can be undesirable because in research that requires intervening on animals and sometimes killing them, more participants may mean more deaths. There can therefore be significant obstacles in the science of consciousness to achieve a β of 0.20 through large samples. Combined, these points suggest that many studies of consciousness will have the standard α of 0.05 (or lower) and β of 0.20 (or higher) because of conventional, resource, or ethical constraints.

5. The conventional standards treat false positives as worse than false negatives

As noted, it is a common ethical intuition among consciousness researchers that false negatives in consciousness ascriptions are prima facie more problematic than false positives. We can now return to the question of whether studies on the dispositions or capacities associated with consciousness generally capture this view, or whether they are normatively neutral. Since many of these studies will for the reasons just outlined have the conventional evidential standards, the answer is negative. This is because the Type II error rate (i.e., β) in these studies is then set to 20%, and the Type I error rate (α) is set to 5%. This means that there is "an implicit asymmetry in the relative importance ascribed to the two types of error. With Type-II error at 20%, this is four times as high as the Type-I error" (Burt et al., 2017, p. 474).

[3] https://www.apa.org/pubs/journals/cns

What is the basis for making the Type I error threshold more stringent than the Type II error threshold? Cohen (1988) writes that the notion that "failure to find is less serious than finding something that is not there accords with (a) the conventional scientific view" and (b) the idea that if we retain the status quo or the default assumption at least we are not exacerbating the situation, i.e., we are not introducing new false claims (p. 56). This matters because it is well known that there is a "publication bias" in science – statistically significant findings are more likely to get published (Martin & Clarke, 2017) – which means that detections and corrections of false positives, as null findings (i.e., replication failures), are less likely to be published. False positive claims may thus become canonized, potentially leading to a significant waste of resources (e.g., if other researchers rely on them). Given these problems and the recent replication failures across many sciences, a number of theorists now argue that the conventional α of 0.05 is too lenient, resulting in too many false positives, and calls for the wide adoption of a more conservative α of 0.005 or 0.001 are increasing (Johnson, 2013; Benjamin et al., 2018; Wasserstein et al., 2019).

In principle, researchers could set their β as low as their α, thus keeping the Type I and II error risks equally low. However, reducing β to even only the 5% level to match the conventional α would require increasing a study's typical power of 80% to 95%, which in turn would require much larger samples than commonly used in natural and social science studies. Systematic analyses found that these studies commonly had only a power of < 50%, suggesting that even the 80% power to keep β at 0.20 (while holding α fixed) may be rarely achieved in many studies (Button et al., 2013; Munafò et al., 2017).

Proposals to reduce α to 0.005 or lower can exacerbate the error asymmetry because α and β are inversely related: if α is reduced (and sample size remains fixed), β increases (Sullivan, 2018). Hence, if researchers reduce α to 0.005, they will need significantly larger samples to even approach the conventional β of 0.20. However, as noted, in many cases, consciousness researchers may not have the resources, or ethical considerations (e.g., to reduce harm to animals) may prevent them from sampling more participants. Of course, since α and β are inversely related, β may also be lowered without a sample size increase by increasing α. But this contradicts the received view and publication norms that require α to be 0.05 or lower.

These intertwined constraints mean that many consciousness researchers may need to set their α to 5% or lower and their β to 20% or higher. This has the consequence that (within the classical statistics framework) it can become inevitable for researchers to accept the value judgment that false positives are more undesirable than false negatives. Yet, this value judgment is the direct opposite to the ethical intuition that many consciousness researchers have expressed, and, as I will argue next, there are good reasons to question it.

6. Problems with the conventional evidential standards in consciousness studies

Focusing on work on animal consciousness, suppose a group of researchers wants to test whether bees display trace conditioning, an ability often taken to be a marker of consciousness (Shea & Bayne, 2010; Birch, 2022). Suppose the researchers detect a difference in their experimental group of bees but their *p*-value is 0.08. Adhering to the current standard α and β of 0.05 and 0.20, respectively, they would conclude that their result is non-significant. However, the researchers would then run a four times higher risk of concluding that there is no evidence that bees are conscious when they are conscious than of concluding that there is evidence that bees are conscious when they are not. The first risk is potentially more harmful for the animals and so potentially more morally problematic, providing consciousness researchers with grounds to question the appropriateness of the conventional evidential benchmarks in consciousness science.

To further illustrate this, consider an example about pain experience. Appel and Elwood (2009) tested whether hermit crabs would meet one key criterion of having pain experience, namely whether they would trade off their response to pain against other motivational requirements. They gave crabs, housed in either preferred or un-preferred shells, electric shocks of increasing intensity to their abdomens to see if they would abandon the shells. Setting α to 0.05 and keeping β at 0.20, Appel and Elwood found that, as "predicted if pain is involved, hermit crabs in preferred shells left the shells at significantly higher voltages than those in un-preferred shells", where "$p = 0.0465$" (p. 122). Appel and Elwood thus conclude suggesting that "pain is felt by crustaceans" (2009, p. 124). But suppose they had instead found a difference with $p = 0.06$. Given their α, they would have concluded that there is no evidence that crabs display a key marker of pain experience, i.e., motivational trade-offs. In doing so, however, Appel and Elwood would have run a much higher risk of missing evidence that crabs can feel pain than of mistakenly concluding that crabs are able to feel pain when they are not. The standard α of 0.05 would therefore again have forced researchers to a morally problematic error trade-off.

While the preceding examples are fictional, there are real cases in which the conventional α and β in fact routinely lead consciousness researchers to swifter postulations of an absence of consciousness than a presence of it. Consider again studies on implicit learning. As mentioned, in these studies, researchers typically aim to prime some form of behavior by a shortly flashed stimulus and then conduct awareness checks in which they assess whether participants fail to detect the stimulus, for instance, in a subsequent recognition task. Systematic reviews found that for these awareness checks, many researchers routinely rely on H0 testing, set α to 0.05, use underpowered samples, and infer the absence of consciousness from null results in discrimination task (Vadillo et al., 2016). Hence, these researchers routinely accept four (or more) times higher risks of false ascriptions of unconscious processing than of false ascriptions of conscious processing.

To clarify, this asymmetry is not per se problematic. It is primarily only problematic in the outlined way when the H0 is that the subject is not conscious or not conscious of *p*. If the H0 instead were that the subject *is* conscious or conscious of *p*, then the risk balance tied to the conventional α and β may align with many consciousness researchers' ethical intuition that in their domain false negatives are more undesirable. For these α and β levels would then mean that mistakenly rejecting that the subject is conscious is treated as four (or more) times more problematic than mistakenly accepting this thesis. Currently, however, most consciousness researchers' default null is not that the subject is conscious or conscious of *p* but the opposite notion (Dienes, 2015; Vadillo et al., 2016).

Yet, there are not only ethical reasons to question the current default H0 of the absence of consciousness and the value judgment underlying the conventional α but also epistemic ones. This is because once statistical relationships are discovered, more studies can, and often do, follow that may confirm, build on, or challenge the original findings such that false positives are uncovered in subsequent research. While publication bias may interfere, the exposure of the findings to social criticism enabled by publication can mitigate the problematic influence of false positives. In contrast, a false negative commonly means that a hypothesis is discarded and no longer tested by the research team because most journals incentivize the report of statistically significant ($p < 0.05$) findings. While the same incentive can also hinder the correction of false positives, the process of scientific self-correction still holds more directly and strongly for false positives than for false negatives, because the former will be more available for social criticism than the latter. The correction of the second kind of error is thus harder in the sense that, unlike false positives, false negatives generally do not even enter the domain of potential social criticism in the first place. As a result, the set of truths discovered in a given scientific domain is smaller than it would be if false positives and false negatives had an equal chance of being followed up (Fiedler et al., 2012).

These points are particularly relevant with regard to studies that explore potential signs of consciousness, because discovering such indicators is especially urgent to be able to mitigate the potential harm of overlooking consciousness in challenging cases. Hindering the discovery of ways of drawing the line between experiencing beings and others is thus vital to avoid. Moreover, it is often noted that the problem of conscious experience, including the detection of the presence of consciousness in challenging cases, is particularly hard and that therefore "stimulating thought and letting a thousand flowers bloom" on these issues, i.e., facilitating discovery, is particularly important (Chalmers, 1994, p. 1). Yet, the standard α of 0.05 in consciousness studies that investigate potential indicators of consciousness does the opposite by privileging the avoidance of false positives over the avoidance of false negatives. And again the related ethical and epistemic risks are significantly increased if, consistent with recent proposals, α is reduced to, for instance, 0.005 across studies, because researchers' false positive chances will then (assuming $\beta = 0.20$) increase from four times as high as the false negative chance to 40 times as severe. That is, recent efforts to mitigate replication failures may push consciousness researchers further towards oversights of consciousness effects than the current evidential standards already do.

To emphasize, the use of the conventional α and β levels or an α reductions in consciousness studies may be warranted. No doubt the proliferation of false positives is highly problematic as it can make the identification of true positives more difficult, and false claims are particularly hard to rectify due to publication bias. Additionally, potential suggestions of weakening the current scientific standards do come with risk of undermining people's trust in the objectivity of consciousness studies. The point here is not to deny or downplay these issues.

The point is rather that it is currently unclear whether the ethical and epistemic costs of false positives in this research are higher than those of an increased number of false negatives. False negatives mean that more areas of consciousness research will remain unexplored and unknown, which is a significant epistemic cost intertwined with ethical risks, given the urgency of finding ways of detecting consciousness in challenging cases. The problem is that the respective costs and benefits of these two kinds of errors in consciousness research have not yet been explicitly compared, but the conventional evidential standards presuppose the value judgment that false positives are more problematic in this domain. That is, as they stand, the current evidential standards in consciousness research rest on a value judgment that has not yet been explicitly justified. Such a justification is needed, however, because without it, scientific studies of consciousness can appear negatively biased against the detection of consciousness effects.

Unfortunately, while there have been important discussions on increased false negative risks related to a rigid adherence to the conventional α in consciousness studies (Cruse et al., 2013; Claassen et al., 2019; Birch, 2023), the specific value judgment underlying the conventional α and β levels have gone largely unnoticed in consciousness research.[4] For instance, while many contributions on the precautionary principle already emphasize the need to take special care to avoid false negatives in consciousness ascriptions (Birch, 2017; Niikawa et al., 2022), they have not yet considered that the current evidential standards underlying many consciousness studies may be much more concerned about controlling false positives than false negatives.

It might be that these standards have so far not been questioned in this debate because theorists in this domain trust that consciousness scientists will have reflect on what the most appropriate error trade-off and benchmarks for statistical significance are. However, this trust could be misplaced because it may also be that many consciousness researchers use classical statistical analyses in their studies and make claims about significance based on $\alpha = 0.05$ even if they have not thought about or do not fully understand the false positives and false negatives balance that this threshold involves (Wasserstein et al., 2019). They may simply follow the convention in their field, and the normative basis may no longer be clearly visible to them, as it has become shared background. The appropriateness of the current evidential standards in consciousness science can therefore not be taken for granted but should be subject to scrutiny, as consciousness researchers may otherwise unwittingly endorse a value judgment that lacks a sufficient justification, is misaligned with their ethical intuitions and hinders the discovery of consciousness indicators.

8. Mitigating the risks

Having argued that the use of the conventional α and β levels in consciousness studies is currently not sufficiently justified, what can be done about the outlined risks related to the error asymmetry captured in these evidential standards? I will discuss four approaches.

A first strategy is to try to increase, whenever feasible and ethically permissible, the sample sizes of consciousness studies. This increases study power, which reduces false negative risks while also mitigating false positives (Button et al., 2013). When ethical considerations or other constraints limit the recruitment of more participants, for instance, multi-laboratory collaborations or within-subject testing may be adopted

[4] Relatedly, in their comprehensive and systematic treatment on the role of ethical values in the science of consciousness, Mazor et al. (2023) do not mention the value judgment underlying current evidential standards in consciousness studies.

(Machery, 2021). However, increasing the sample size to the point that the risks of false positives and false negatives are fully equalized may often not be realistic due to resource limitations.

Another, albeit more problematic approach to treating false positives and false negatives more equally would be to relax α, which, as noted, can reduce false negatives. False positives and replication failures of studies on consciousness will then increase too, which is undesirable. However, even advocates of the view that the conventional α should be reduced to 0.005 tend to acknowledge that we "should examine how easily the sample size can be increased and what the respective costs of false positives and false negatives are. On this basis, scientists should carefully set their alpha level at a particular, context-sensitive value, and no conventional alpha level is needed" (Machery, 2021, p. 93). Indeed, the current "blanket α" of 0.05 across scientific fields, combined with β of 0.20, requires researchers to "pretend that the relative importance of Type I and Type II errors is constant" across studies even though it is not (Trafimow et al., 2018, p. 3). To the extent that scientists should be allowed to set their α to a context-sensitive value, it might seem consciousness researchers, too, should be allowed to relax their α to accommodate the high costs of false negatives if sample size increases are not feasible. To prevent that they specify α so that any of their results appear significant, α could be set institutionally and differently in specific areas (Machery, 2021). Moreover, if there is an increased, institutionally agreed and encouraged openness in the field of consciousness science toward publishing negative results, this may buffer an increase in false positives that could follow a relaxation of the conventional α, as publishing negative findings can facilitate corrections of these errors. Weaking α in this domain may therefore be less problematic than it might seem, provided it happens simultaneously with such a social and institutional change. However, this approach remains unattractive given the risk of undermining public trust in consciousness science (Birch, 2017).

A third, more appealing strategy is to encourage researchers to use *p*-values as a measure of evidence against the H0 without referring to a specific significance cut-off point (Hurlbert & Lombardi, 2009). No dichotomous claim on whether there is an effect would then be made based on *p*-values, and so false positives and false negatives would not need to be traded off. However, especially when it comes to the presence or absence of consciousness effects, policymakers and researchers alike may often need a dichotomous verdict to make informed decisions, including on whether to accept a particular behavioral, cognitive, or neural feature as a consciousness indicator (Farisco et al., 2022). Another problem is that while *p*-values can be a measure of evidence against the H0, they do not provide direct evidence for the H1; nor do they provide evidence for accepting the H0 since a result might be non-significant because the study was not sensitive enough to detect effects (Dienes, 2015).

To address this problem, it has been recommended that consciousness researchers avoid the *p*-value and instead calculate the Bayes Factor (BF) of results (Dienes, 2021). The BF is the ratio of the likelihood of the observed data under the H0 to the likelihood of the data under the alternative hypothesis ('H1') (Kruschke & Liddell, 2018). It allows quantifying the strength of evidence for both the H0 and H1. For example, while a BF of 1 indicates that the data is equally likely under either H0 or H1, a $BF_{10} = 4$ means that, given the data, the evidence for the H1 is four times stronger than for the H0. A $BF_{10} = 1/4$ indicates the reverse. Hence, while non-significant *p*-values in, for instance, consciousness checks in implicit learning studies do not provide evidence that the processing was unconscious, BFs can do so, helping to avoid conflating the absence of evidence of consciousness with evidence of absence of consciousness (Dienes, 2015). Relatedly, while non-significant *p*-values are commonly used to conclude that there is *no* evidence that the H1 is true, BFs instead specify the strength of the evidence for the H1, facilitating more informed decision-making.

The use of BFs has further advantages. Compared to the frequentist approach, Bayesian hypothesis testing can allow researchers to stop recruiting participants and report the findings of a hypothesis test when the data already indicate compelling evidence after only a part of the intended participant number is assessed (Hoijtink et al., 2019; Bendtsen, 2022). This is a significant benefit especially for consciousness researchers because a lack of resources, or ethical constraints may often restrict their ability to recruit large samples. And crucially, using BFs can also allow consciousness researchers to avoid the judgment that false positives are more undesirable than false negatives, because BFs alone are not dichotomous but continuous, presenting evidence of different strengths for or against the H0 or H1 (Held & Ott, 2018).

However, there are limitations too. Calculating the BF requires specifying the prior probability of the H1 and H0. This introduces a subjective choice and comes again with inductive risks because holding that the prior probability of the H1 is X means accepting a hypothesis H* about the probability of H1. Bayesians can therefore not avoid the need to settle whether evidence is sufficiently strong to support accepting a hypothesis. In Bayesian testing, it is simply a different hypothesis with respect to which one now needs to choose, namely H* (Parker & Winsberg, 2018). Moreover, even if the subjective influence on setting the priors can be reduced (e.g., by using default priors; Halsey, 2019), since BFs alone leave it unclear whether a hypothesis should be accepted, BFs alone may (just as *p*-values without α levels) only be of limited use, as researchers, reviewers, or editors need a justifiable criterion to decide about research publication and future studies. Bayesians therefore often employ conventional values such as a BF of > 3, or > 10 to make decisions and dichotomous claims (Schmalz et al., 2021, p. 11). Yet, when Bayesians go beyond simply informing researchers how much they should revise their belief in a hypothesis and produce dichotomous verdicts, their claims can be mistaken, which re-introduces the need to determine the false positive and false negative rates connected with a BF of a certain value. And importantly, simulation studies focusing on Bayesian versions of the most common frequentist two-sample tests found that with commonly used priors and a $BF_{10} > 3$ threshold, “Bayesian two-sample tests yield better type I error control at the cost of slightly increased type II error control compared to their frequentist counterparts”; for instance, for small effects, “Bayesian tests need a larger sample size to achieve the same type II error rate (the same power) as the frequentist two-sample tests” (Kelter, 2021, p. 1284). That is, even if researchers use Bayesian tests in consciousness studies, to the extent that they also invoke common decision thresholds to make dichotomous claims, they may exacerbate the problematic error balance captured in the value judgment underlying the conventional α and β levels.

Nonetheless, replacing frequentist testing with Bayesian testing remains an attractive strategy for consciousness researchers to mitigate the risks related to this value judgment, because dichotomous claims may not always be needed in consciousness studies, and when they are not required, the quantification of evidence for both H0 or

H1 that BF provide is more informative than *p*-values for the reasons outlined above (Halsey, 2019). Additionally, when using Bayesian statistics, specific decision thresholds (e.g., BF > 3) are currently less established than the conventional α and β levels because the use of Bayesian testing is still not very common across the sciences (Schmalz et al., 2021). This can make it significantly easier for consciousness researchers who use BFs instead of *p*-values to adopt a decision threshold with more balanced false positive and false negative rates than the conventional benchmarks. Finally, unlike the use of *p*-values, which is often done unreflectively (Wasserstein et al., 2019), to use BFs, researchers first need to specify their priors, which requires an understanding and acknowledgment of the implicit assumptions that the prior makes about the studied system, how they shape the posterior probabilities, and transparent communication to potential recipient of the results (Banner et al., 2020). The use of BFs may therefore promote transparency in consciousness studies about the underlying assumptions and prompt researchers to explore alternative inductive risk judgments than the one currently underlying the conventional statistical benchmarks.

These points do not only suggest that Bayesian testing is a promising approach to better align the currently used benchmarks with the common ethical intuition that false negatives are especially undesirable in this domain. They in fact also reversely provide the basis for a new argument for adopting Bayesian testing in science, specifically, in consciousness science: Replacing *p*-values and frequentist testing with BFs and Bayesian testing in experimental research on consciousness allows mitigating the ethical and epistemic risks related to the conventional evidential standards (i.e., the increased chances of oversight of consciousness or responsiveness when it is present). Hence, while there are already persuasive epistemic rationales for the Bayesian approach in science, in general (Schmalz et al., 2021), and consciousness research (including medical contexts), in particular (Dienes, 2021; Birch, 2023), this paper adds an ethical one connected to the error balance underlying current evidential standards.[5]

9. Conclusion

Finding ways of detecting consciousness in challenging cases (e.g., comatose patients) is urgent. Scientific studies that investigate and identify signs of consciousness play a key role in this effort. Here, I examined whether these studies capture the common ethical intuition that false negatives in consciousness ascriptions are prima facie more problematic than false positives. I argued that for many of these studies this is not the case because they rely on the same evidential standards as other scientific studies, setting their α and β to 0.05 (or lower) and 0.20 (or higher), respectively. I noted that these thresholds rest on the value judgment that false positives are much more undesirable than false negatives, and this value judgment may (inter alia) hinder the discovery of signs of consciousness. Recent attempts to make α more stringent might exacerbate these problems because consciousness researchers may often not be able to

[5] Focusing on medical contexts, Birch (2023) also argues that the use of Bayesian statistics can have ethical benefits. Specifically, he suggests that BF bounds (Benjamin & Berger, 2019) combined with odds upper bounds and probability yardsticks can help clinicians provide a vegetative patient's family with more relevant information about the patient's potential responsiveness (or consciousness) than binary verdicts of 'responding'/'not responding' or false discovery rates. The argument here is congenial to but different from this point. It is that the adoption of a Bayesian approach is ethically beneficial because it can allow consciousness researchers to avoid evidential standards that treat oversights of consciousness as less problematic than over-ascriptions of consciousness.

recruit larger samples to prevent the corresponding inflation of false negative rates. While the use of the conventional evidential standards in consciousness studies might nonetheless be justified (e.g., by the higher costs of false positives), the main point here was that these costs have not yet been weighed against the costs of false negatives in consciousness research. Hence, using the conventional α and β levels in this research is currently insufficiently justified, which may encourage policymakers, the public, and researchers to view consciousness studies as unduly negatively skewed against the detection of consciousness effects. To mitigate this, I recommended replacing the conventional significance benchmarks in consciousness science with Bayes Factors. I noted that the advantages that the Bayesian approach offers in reducing the ethical and epistemic risks related to the conventional benchmarks in this domain speak themselves in favor of adopting this approach in science, specifically, in studies of consciousness. Hence, by making explicit a key value judgment underlying standard frequentist significance thresholds and by highlighting that this value judgment is at odds with a common ethical intuition and insufficiently justified, this paper offers new support for an increasingly stronger case for a Bayesian turn in the science of consciousness.

References

Appel, M., & Elwood, R. W. (2009). Motivational trade-offs and potential pain experience in hermit crabs. *Applied Animal Behaviour Science*, 119(1), 120–124.

Banner, Katharine & Irvine, Kathryn & Rodhouse, Thomas. (2020). The Use of Bayesian Priors in Ecology: The Good, The Bad, and The Not Great. Methods in Ecology and Evolution. 11. 10.1111/2041-210X.13407.

Bayne, T., & Shea, N. (2021). Consciousness, concepts, and natural kinds. *Philosophical Topics*, 48(1), 65–83.

Bendtsen M. (2022). Avoiding Under- and Overrecruitment in Behavioral Intervention Trials Using Bayesian Sequential Designs: Tutorial. *Journal of medical Internet research*, *24*(12), e40730. https://doi.org/10.2196/40730

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., … Johnson, V. E. (2018). Redefine statistical significance. *Nature human behaviour*, 2(1), 6–10.

Benjamin, D.J. & Berger, J.O. (2019) Three Recommendations for Improving the Use of *p*-Values. *The American Statistician*, 73, 1, 186
191, DOI: 10.1080/00031305.2018.1543135

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, 2 (16).

Birch, J. (2022). The search for invertebrate consciousness. *Noûs*_

Birch, J. (2023). Medical AI, Inductive Risk, and the Communication of Uncertainty: The Case of Disorders of Consciousness. *Journal for Medical Ethics.* https://eprints.lse.ac.uk/120570/1/Inductive_risk.pdf

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18, 227–247.

Bradshaw, R. H. (1998). Consciousness in non-human animals: adopting the precautionary principle. *Journal of Consciousness Studies* 5:108-114.

Brown, C. (2016). Fish pain: An inconvenient truth. *Animal Sentience* 3(32).

Brown, D. & Key, B. (2019). You look but do not find: why the absence of evidence can be a useful thing. *The Conversation*. URL: https://theconversation.com/you-look-but-do-not-find-why-the-absence-of-evidence-can-be-a-useful-thing-114988

Burt, T., Button, K. S., Thom, H., Noveck, R. J., & Munafò, M. R. (2017). The Burden of the 'False-Negatives' in Clinical Development: Analyses of Current and Alternative

Scenarios and Corrective Measures. *Clinical and translational science*, 10(6), 470–479.

Chalmers, D. J. (1994). Review of Journal of Consciousness Studies. *Times Literary Supplement*, URL: https://consc.net/papers/tls.html

Chalmers, D.J. (2013), How can we construct a science of consciousness?. *Ann. N.Y. Acad. Sci*., 1303: 25-35.

Claassen, J., Doyle, K., Matory, A., Couch, C., Burger, K. M., Velazquez, A., Okonkwo, J. U., King, J. R., Park, S., Agarwal, S., Roh, D., Megjhani, M., Eliseyev, A., Connolly, E. S., & Rohaut, B. (2019). Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury. *The New England journal of medicine*, *380*(26), 2497–2505.

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

Clark, R. E., & Squire, L. R. (1999). Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science*, 10, 14–18.

Cleeremans, A. & Tallon-Baudry, C. (2022). Consciousness matters: phenomenal experience has functional value. *Neuroscience of Consciousness*, 1.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T. A., Fernández-Espejo, D., Pickard, J. D., Laureys, S., & Owen, A. M. (2011). Bedside detection of awareness in the vegetative state: a cohort study. *Lancet (London, England)*, *378*(9809), 2088–2094.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T. A., Fernández-Espejo, D., Pickard, J. D., Laureys, S., & Owen, A. M. (2013). Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study" - Authors' reply. *Lancet (London, England)*, *381*(9863), 291–292.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395, 597–600.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (pp. 199–220). Oxford: Oxford University Press.

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9–26. https://doi.org/10.1037/cns0000258

Di Leo, G., & Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics-reasons for a conservative approach. *European radiology experimental*, 4(1), 18.

Douglas, H. E. (2009). *Science, Policy, and the Value-free Ideal*. University of Pittsburgh Press.

Dung, L. & Newen, A.(2023). Profiles of animal consciousness: A species-sensitive, two-tier account to quality and distribution. *Cognition* 235 (C):105409.

Elliott, K. C., & Richards, T. (2017). *Exploring Inductive Risk*. Oxford University Press.

Farisco, M., Pennartz, C., Annen, J. *et al.* Indicators and criteria of consciousness: ethical implications for the care of behaviourally unresponsive patients. *BMC Med Ethics* **23**, 30 (2022). https://doi.org/10.1186/s12910-022-00770-3

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The Long Way From α-Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate. *Perspectives on Psychological Science*, 7(6), 661–669.

Fins, J. J., & Bernat, J. L. (2018). Ethical, palliative, and policy considerations in disorders of consciousness. *Neurology*, *91*(10), 471–475.

Halsey L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?. *Biology letters*, *15*(5), 20190174. https://doi.org/10.1098/rsbl.2019.0174

Held, L. & Ott, M. (2018). On p-Values and Bayes Factors. *Annual Review of Statistics and Its Application*, 5(1):593-419.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556.

Ioannidis, J.P.A. (2019) What Have We (Not) Learnt from Millions of Scientific Papers with P-Values? *The American Statistician*, 73, 1, 20-25.

Johnson V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48), 19313–19317.

Jones, R. C. (2016). Fish sentience and the precautionary principle. *Animal Sentience* 3(10).

Kelter, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*. 36. 1-26. 10.1007/s00180-020-01034-7.

Koplin, J.J., & J. Savulescu (2019). Moral Limits of Brain Organoid Research. *The Journal of Law, Medicine & Ethics* 47, 760–767.

Kruschke, J.K., & Liddell, T.M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review,* 25:155–177.

Lakens, D., Adolfi, F.G., Albers, C.J. et al. (2018). Justify your alpha. *Nat Hum Behav* 2, 168–171.

Levy N. (2014). The Value of Consciousness. *Journal of consciousness studies*, 21(1-2), 127–138.

Machery, E. (2021). The Alpha War. *Rev.Phil.Psych*. 12, 75–99. https://doi.org/10.1007/s13164-019-00440-1

Magnus, P.D. (2022). The scope of inductive risk. *Metaphilosophy*, 53(1): 17–24.

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, 8, 523.

Mazor, M., Brown, S., Ciaunica, A., Demertzi, A., Fahrenfort, J., Faivre, N., Francken, J. C., Lamy, D., Lenggenhager, B., Moutoussis, M., Nizzi, M. C., Salomon, R., Soto, D., Stein, T., & Lubianiker, N. (2023). The Scientific Study of Consciousness Cannot and Should Not Be Morally Neutral. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 18(3), 535–543.

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PloS* one, 7(2), e32734.

Munafò, M., Nosek, B., Bishop, D. et al. (2017). A manifesto for reproducible science. *Nat Hum Behav* 1, 0021.

Niikawa, T., Hayashi, Y., Shepherd, J., & Sawai, T. (2022). Human brain organoidsand consciousness. *Neuroethics*, 15(1), 1–16.

Nuzzo R. (2014). Scientific method: statistical errors. *Nature*, 506(7487), 150–152.

Katz M. H. (2006) *Study design and statistical analysis: A practical guide for clinicians.* Cambridge, UK: Cambridge Univer. Press.

O'Riordan, T. T Cameron, J. (Eds.). (1994). *Interpreting the precautionary principle.* London: Earthscan.

Parker, W.S. & Winsberg, E. (2018). Values and evidence: how models make a difference. *Euro Jnl Phil Sci* **8**, 125–142, https://doi.org/10.1007/s13194-017-0180-6

Perruchet, P. (1985). Expectancy for Airpuff and Conditioned Eyeblinks in Humans. *Acta Psychologica*, 58, 31–44.

Peterson, A., Cruse, D., Naci, L., Weijer, C., & Owen, A. M. (2015). Risk, diagnostic error, and the clinical science of consciousness. *NeuroImage. Clinical*, *7*, 588–597. https://doi.org/10.1016/j.nicl.2015.02.008

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M. A., Mariotti, M., Boveroux, P., Tononi, G., Laureys, S., & Massimini, M. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain : a journal of neurology*, *135*(Pt 4), 1308–1320. https://doi.org/10.1093/brain/awr340

Seth, A. K. (2016). Why fish pain cannot and should not be ruled out. *Animal Sentience* 3(14).

Shea, N., & Bayne, T. (2010). The vegetative state and the science of consciousness. *British Journal for the Philosophy of Science*, 61, 459–484.

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Methods, 28*(3), 705 718. https://doi.org/10.1037/met0000421

Strawson, G. (1994). *Mental Reality*. Cambridge, Mass.: MIT Press (1994)

Sullivan M. (2018) *Statistics: Informed decisions using data*. (2nd ed.) Upper Saddle River, NJ: Pearson/Prentice Hall.

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., Bono, R., Bradley, M. T., Briggs, W. M., Cepeda-Freyre, H. A., Chaigneau, S. E., Ciocca, D. R., Correa, J. C., Cousineau, D., de Boer, M. R., Dhar, S. S., Dolgov, I., Gómez-Benito, J., Grendar, M., Grice, J. W., … Marmolejo-Ramos, F. (2018). Manipulating the Alpha Level Cannot Cure Significance Testing. *Frontiers in psychology*, 9, 699.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic bulletin & review*, *23*(1), 87–102.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European journal of psychotraumatology*, *6*, 25216. https://doi.org/10.3402/ejpt.v6.25216

Vidgen, B., & Yasseri, T. (2016). P-values: misunderstood and misused. *Frontiers in Physics*, 4.

Wasserstein, R., Schirm, A.L., & Lazar, N.A. (2019). Moving to a World Beyond ‘p < 0.05’. *The American Statistician*, 73, 1, 1-19.