Calibration in Consciousness Science

Matthias Michel[1]

1. Bersoff Faculty Fellow, New York University

Abstract: To study consciousness, scientists need to determine when participants are conscious and when they are not. They do so with consciousness detection procedures. A recurring skeptical argument against those procedures is that they cannot be calibrated: there is no way to make sure that detection outcomes are accurate. In this article, I address two main skeptical arguments purporting to show that consciousness scientists cannot calibrate detection procedures. I conclude that there is nothing wrong with calibration in consciousness science.

Introduction

To test hypotheses about consciousness, scientists need to determine whether subjects are conscious of stimuli or not[1]. They do so by using methods for eliciting behaviors that can be analyzed and interpreted as indications of consciousness. I call these methods *consciousness detection procedures*[2] (Irvine, 2013; Spener, forthcoming; Timmermans & Cleeremans, 2015). To be of any use, these procedures need to be reliable. Calibration is the process by which scientists evaluate the reliability of detection procedures, and correct them if necessary.

Skeptics, as Irvine (2012b, 2019), Feest (2012, 2014), Goldman (2002), and Spener (2013, 2015) hold that detection procedures based on the subjects' introspective reports cannot be calibrated[3]. Calibration requires several valid ways of detecting the same phenomenon. But researchers claim that introspection-based procedures are the only valid way of detecting consciousness. So, the argument goes, calibration is out of reach in consciousness science.

One alternative to calibrate detection procedures is to learn more about introspection and identify the factors influencing subjective reports, as Spener (2015, p.13-14) suggests. But Irvine (2012, 2019) and Schwitzgebel (2012) hold that introspection and the factors that could influence it are too complex to be successfully modelled. Introspection is "a cognitive confluence of crazy spaghetti" (Schwitzgebel, 2011, p.41).

This argument justifies radical skepticism towards the calibration of consciousness detection procedures. If the skeptics are right, consciousness science is in trouble. Having a respectable science of consciousness requires valid and reliable detection procedures. But there seems to be no way of knowing whether those procedures are reliable or not. The scientific

---

[1] A subject is conscious of a stimulus if there's "something it's like" for her to perceive it (Nagel, 1974). She perceives a stimulus unconsciously if she mentally represents that there is a stimulus, but doing so does not feel any different from not representing that there's a stimulus.

[2] Consciousness scientists and philosophers alike often talk about "measures" of consciousness (Irvine, 2013; Timmermans & Cleeremans, 2015; Sandberg et al., 2010; Seth et al., 2008; Spener, forthcoming). I prefer to talk about the "detection" of consciousness. Measurement is supposed to apply to quantitative properties (Joint Committee for Guides in Metrology, 2012; Michell, 1999). Using the term "measurement" assumes that consciousness is a quantitative property, and that the outcomes of procedures used by consciousness scientists attribute degrees of consciousness of stimuli to the subjects. In most experiments, consciousness scientists simply try to assess whether or not subjects are conscious of stimuli, that is, they attempt to detect consciousness.

[3] Skepticism about the calibration of consciousness detection procedures comes in degrees. For instance, while Spener (2015) accepts some of the skeptics' arguments, she argues that we can identify conditions in which introspective judgments tend to be reliable (see also Bayne & Spener, 2010). Irvine (2012, 2019) is a skeptic in the most straightforward sense, and I mainly focus on her arguments here.

beliefs acquired in consciousness science would thus hang on potentially contentious and unjustified detection procedures (Goldman, 2002, 2004; Michel, 2019; Spener, 2015).

I disagree with the skeptics. Consciousness detection procedures can be successfully calibrated. I begin with a brief introduction to the detection procedures typically used in consciousness science. Then, I present two skeptical arguments: Irvine's "arbitrariness argument" (Irvine, 2012, 2019), and Schwitzgebel's "crazy-spaghetti argument" (Schwitzgebel, 2012). Finally, I answer those arguments, and conclude that there is nothing wrong with calibration in consciousness science.

1. Detection procedures in consciousness science

You are sitting in front of a screen to participate in a typical consciousness science experiment. A stimulus is flashed, quickly followed by a mask. You have to answer whether the stimulus was a square or a diamond by pressing a button. A text then appears on the screen: "Seen or Guessed?" You press a button to give a response. Repeat this about two thousand times, and finally, receive some money for your participation.

You have two tasks in this experiment. The first is called the *Type-1 task*. It requires you to make perceptual decisions about the external world. In the *Type-2 task*, you make decisions about your own mental states (Galvin et al., 2003). In that sense, the Type-2 task is a metacognitive, or introspective task (Chirimuuta, 2014)[4]. Based on your responses, scientists decide whether you were conscious or unconscious of the stimuli. They have procedures for doing this: consciousness detection procedures.

Researchers distinguish between two main types of consciousness detection procedures: Type-1-based and Type-2-based procedures (Irvine, 2013; Timmermans & Cleeremans, 2015; Seth et al., 2008)[5].

---

[4] We should distinguish between metacognition and introspection. As I use the terms, metacognitive representations are about cognitive or perceptual states. Representations that result from introspection, on the other hand, are about the subject herself. In a slogan: metacognition is cognition about cognition, introspection is cognition about the cognizer. While introspection is by definition a metacognitive process, not all metacognitive processes are introspective. For instance, a poker player might monitor her own states of uncertainty, and make use of those states, without having to consciously self-attribute those states (Carruthers & Ritchie, 2012). In this article, I assume that Type-2 tasks are generally introspective in nature because they require subjects to self-attribute perceptual or confidence states.

[5] Philosophers and scientists often draw the distinction in terms of "objective" and "subjective" procedures. The latter are probably called "subjective" because they rely on "subjective" reports. But

In Type-1-based procedures, scientists analyze the participants' Type-1 responses to determine whether they were conscious of the stimuli or not. To do so, they can employ the Signal Detection Theoretic measure of perceptual sensitivity called *d'*. Roughly, *d'* quantifies how effectively a (perceptual) system can discriminate between objective states of the world (Macmillan & Creelman, 2005).

Perceptual sensitivity has a straightforward interpretation in terms of consciousness. If you are able to meaningfully respond to stimuli, you are more likely to be conscious of them than if you're unable to do so. Type-1-based procedures consist in interpreting *d'* as an indication of consciousness of stimuli (Holender, 1986; Cheesman & Merikle, 1986; Snodgrass et al., 2005).

Type-2-based procedures rely on both Type-1 and Type-2 responses to decide whether subjects are conscious of stimuli or not. Type-2 responses can be provided with a variety of scales, such as confidence or visibility scales (Sandberg & Overgaard, 2015; Norman & Price, 2015).

A straightforward way to interpret Type-2 responses is to treat them as direct indications of consciousness. But consciousness detection procedures don't generally work that way. After all, things could go terribly wrong on any given trial. Perhaps on this trial the subject answered too quickly, or wasn't paying attention, or pressed a random button to move on to the next trial because she's starting to get bored, or decided to respond "seen" because she had answered "unseen" on the last five trials and felt like changing, or for whatever reason thought that the experimenter would expect her to answer "seen" on this trial, and so forth. Because any given response could be just a fluke, scientists typically ignore whether any particular Type-2 response is a reliable indication of consciousness (Timmermans & Cleeremans, 2015; Schmidt, 2015).

Partly for this reason, Type-2 responses are not interpreted as indicating consciousness of stimuli on a trial-by-trial basis. Instead, patterns of Type-1 and Type-2 responses are analyzed through the tools of Signal Detection Theory to determine the subjects' consciousness of stimuli throughout a task, or condition of a task. By doing so "one abandons the ability to establish, for any single stimulus, whether it was consciously perceived or not" (Timmermans & Cleeremans, 2015, p.38). In exchange, one virtually eliminates the influence of fluke responses on the detection procedure. I'll come back to this important point later.

---

those reports consist of publicly observable behaviors like pressing buttons, as observed by Piccinini (2009). It is unclear what is "subjective" in pressing buttons, or even in verbal reports. In any case, "Type-1-based" and "Type-2-based" procedures is a more neutral way of putting the distinction.

To interpret and analyze the subjects' responses, researchers often look for correlations between patterns of Type-1 and Type-2 responses[6]. Those correlations are then typically interpreted in light of two criteria: the zero correlation criterion, and the guessing criterion (Dienes et al., 1995;Dienes & Perner, 2004).

The zero correlation criterion says that a positive correlation between a pattern of confidence/visibility ratings and Type-1 accuracy assessed by *d'* should be interpreted as indicating consciousness. The name 'zero correlation criterion' refers to the idea that if a subject perceives a stimulus but is totally unaware of perceiving it, there should be no correlation between Type-1 decisions and Type-2 ratings. On the other hand, a correlation between Type-1 decisions and Type-2 ratings suggests that the observer is able to adjust her Type-2 reports based on the information that she uses during the Type-1 task[7]. And visual information that an observer can use to modulate her Type-2 responses is more likely to be conscious than unconscious (Dienes & Perner, 2004).

The "guessing criterion" says that above-chance Type-1 performance on trials where the subject used the lowest confidence/visibility ratings indicates unconscious perception. The subject was seemingly able to use visual information to successfully perform the Type-1 task, but her Type-2 responses suggest that she was unaware of that (Dienes et al., 1995). This motivates the claim that she was not conscious of some of the sensory information used to perform the Type-1 task[8].

This should be enough for detection procedures. I'll give more details along the way. The point to keep in mind is that there are crucial differences between the various detection

---

[6] I will not discuss the statistical tools that scientists use to find those correlations (for reviews, see Fleming & Lau, 2014; Maniscalco & Lau, 2014). The gold-standard is currently to compute an index called *Meta-d'*, which is then used to quantify how much of the information used during the Type-1 task was available for the subject to perform the Type-2 task (Maniscalco & Lau, 2012).

[7] This correlation could be driven by a third factor influencing both Type-1 and Type-2 decisions in the same way, instead of being explained by the influence of visual information used for the Type-1 task over the Type-2 decision. At this stage, this 'third factor' hypothesis cannot be ruled out, although it is unclear what that third factor would be. The hypothesis that participants take Type-2 decisions based on what they consciously see during the Type-1 task seems to be a simple and straightforward explanation of the correlation between Type-1 and Type-2 performance.

[8] Those two criteria are complementary. There's a big difference between not seeing something and seeing something unconsciously. I don't see what's behind my head consciously, but I don't see it unconsciously either. The zero-correlation criterion allows experimenters to decide whether or not a large proportion of stimuli were consciously perceived during the task. The guessing criterion allows them to decide whether the stimuli that were not perceived consciously were *perceived* unconsciously or not perceived at all.

procedures that consciousness scientists have at their disposal, the most important being the difference between Type-1-based and Type-2-based procedures.

These procedures often provide different outcomes for the same tasks (e.g., Cheesman & Merikle, 1986; Holender, 1986; Li et al., 2014; Overgaard et al., 2006; Rausch et al., 2015; Sandberg et al., 2010; Snodgrass et al., 2005; Szczepanowski et al., 2013; Wierzchoń et al., 2014). So, the question is: which of these procedures, if any, allow scientists to accurately infer whether subjects are conscious of stimuli or not?

Answering this question requires the ability to evaluate these procedures, and correct them if necessary — to *calibrate* consciousness detection procedures[9]. Skeptics have argued that calibration is particularly difficult, if not impossible, to achieve in consciousness science. I now present the two main arguments supporting this claim.

2. Motivating radical skepticism

Here's how Irvine (2012a) summarizes what I call the "arbitrariness argument", which purports to show that scientists cannot calibrate Type-2-based procedures:

> Aside from intuitions about what is likely to be the right way of reporting (which themselves are subject to bias), there is no way of providing a metric for categorizing biases and reports. This is because *introspection is proposed to be the only method* of investigating conscious phenomena. With no other points of reference, even extreme biases cannot be discounted as bad, and introspective reports cannot be evaluated as more or less correct. That is, 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, 2012a, p.634; my emphasis)

Spener (forthcoming) expresses the same worry:

> In relying on subjective reports to provide access to consciousness, subjective measures do not seem open to independent validation in the form of inter-measurement calibration (...) As a

[9] Attempts to calibrate procedures relying on 'subjective' or 'introspective' judgments are not new. For instance, Titchener's *Experimental Psychology: A Manual of Laboratory Practice* (1905) can be considered as an 'introspective training manual' (Schwitzgebel, 2011; Chapter 5), allowing psychologists to improve introspective 'observations' (See Kroker, 2003). With respect to calibration, an important difference with contemporary consciousness science is that calibration in Titchener's practice was focused on the *introspector*, not on the whole detection process, including experimental settings, as well as methods of statistical analyses. As Lyons (1986) puts it, calibration was more concerned with 'the attitude of the introspector' than 'the laboratory conditions themselves' (p.19).

> consequence, such measures have faced sharp criticism for apparent failure to comply with sufficiently rigorous scientific standards. (Spener, forthcoming, p.2)

Calibration requires several methods for detecting or measuring the same phenomenon. Scientists hold that introspection-based methods are the only valid methods for detecting consciousness of stimuli. Therefore, scientists cannot calibrate consciousness detection procedures.

The first premise of this argument is true in the case of "concordance-calibration", which is a specific method for calibrating procedures (I call the other one "model-calibration")[10]. If different procedures relying on different assumptions provide similar outcomes, this counts as evidence in favor of the accuracy of those procedures. Why? Because otherwise systematic agreement between those two widely dissimilar procedures would be a preposterous coincidence (Hacking, 1981)[11]. The best explanation is that both procedures detect the same thing with similar degrees of accuracy. If the outcomes of the two procedures disagree, one of the two procedures, or both, are inaccurate. This kind of "concordance-calibration" allows one to evaluate whether procedures are accurate or inaccurate. What matters for our purpose is that concordance-calibration requires several procedures for detecting the same phenomenon.

Let's look at the second premise now. Blindsight subjects can have above chance performance in Type-1 tasks, and yet, claim that they don't see anything in their blind field (Cowey, 2010; Weiskrantz, 2009). For this reason scientists often hold that Type-1-based procedures are not valid for detecting consciousness (Lau, 2008; LeDoux, 2019; Morales et al., 2015, Weiskrantz, 1998). And if Type-1-based procedures are not regarded as valid, the skeptics argue, scientists cannot calibrate Type-2-based procedures by comparing their outcomes with those of Type-1-based procedures (Irvine, 2019).

---

[10] This distinction is inspired from Tal's distinction between black-box and white-box calibration of measuring instruments (Tal, 2017).

[11] I assume that concordance-calibration can provide a good indication of accuracy, if not a sufficient condition for establishing accuracy. This point has been debated (Basso, 2017). Most researchers recognize the value of comparing various measurement procedures, while disagreeing on the reason *why* doing so is valuable. Concordance-calibration could be useful because it allows one to compare procedures that do not share the same sources of error, namely, *independent* procedures (Kuorikoski et al. 2012). An alternative source of value for concordance-calibration is to be found, not in the *independence* of measurement procedures, but in their *complementarity*: different procedures have different strengths and weaknesses. As Basso (2017) puts it: "Since each procedure can fail to realize the definition (and hence to measure the quantity of interest) in different ways, the convergence of their results can be taken as a sign that these sources of uncertainty do not lead the results completely astray and, therefore, that the procedures measure the quantity as defined with sufficient reliability." (p.8).

If you buy these two premises then concordance-calibration is out of reach for consciousness science. If so, preferring a given procedure to a different procedure that would provide potentially different outcomes is arbitrary.

The arbitrariness argument alone is not enough to warrant general skepticism toward the calibration of consciousness detection procedures. It only targets one specific form of calibration: concordance-calibration. What I call "model-calibration" does not require different ways of detecting the same phenomenon. Let me explain how it works.

Model-calibration starts by learning more about the procedure itself. By developing a model of the way in which the procedure works, you can attempt to predict the indications that the procedure *should* output, in a given situation, *if it were accurate*. You can then compare the actual outcome of the procedure with the outcomes predicted by the model to calibrate it.

In addition, if you suspect that a confounding factor C influences the procedure, you can create a model of the way in which the procedure would work *if it were influenced by C*. You now have a model of the procedure as it should *ideally* work, and a model of the procedure as it would work if it were influenced by C. If the actual indications produced by the procedure fit the predictions of the latter model, this is a good reason to believe that C is a confounding factor. You can then recalibrate the procedure such as to minimize the influence of C.

This method was used, for instance, in the calibration of cesium fountain clocks (Tal, 2014, 2017). Scientists made theoretical assumptions about the effects of atomic collisions and magnetic fields on the indications of the clock to model what the indications *would have been* without these effects. They then compared the actual indications of the clock to those it would have provided in different conditions to quantify the effects of potential confounding factors on measurement indications. In this case, successful calibration was achieved by building a *model* of the procedure itself.

Irvine (2019) and Spener (2015) recognize that model-calibration is an alternative to the kind of concordance-calibration ruled out by the arbitrariness argument. Irvine writes that an obvious way to solve the calibration problem is "to just learn more about the process of introspection" (p.20):

> If we knew more about how introspective reports are generated, we could gain more confidence in developing and applying introspective procedures, and so improve the evidentiary value of introspective reports (Irvine, 2019, p.20).

To motivate radical skepticism against the possibility of calibrating consciousness detection procedures, skeptics thus need to provide an argument against model-calibration.

Enter the "crazy-spaghetti argument". Irvine writes:

> One could imagine ... [trying] to isolate and manipulate factors that are traditionally out of bounds of psychophysics (...) these might include self-shaping or self-fulfilment, memory and association. Once these factors are mapped out, one might then be able to control for them experimentally, or otherwise try to interpret introspective reports in light of them. This would not be easy: the number of factors that might be relevant, and how they affect introspective reports in isolation and in combination, is likely to be complex, to vary significantly across tasks, and also likely to vary across individuals (certainly for the more cognitive factors). At the very least, this will take a while. (Irvine, 2019, p.21)

Model-calibration requires scientists to build a model of the procedure itself to determine what the indications (i.e., introspective reports) *would have been* in different conditions. But a wide variety of factors could influence introspective reports in (yet) unknown ways. As Schwitzgebel writes, introspection looks like "a cognitive confluence of crazy spaghetti" (2011, p.41). And since we ignore how those factors influence the detection outcomes, we can't build a model of the detection procedure. Therefore, we can't use model-calibration in consciousness science.

With the arbitrariness and crazy-spaghetti arguments combined, the skeptics can rule out both concordance- and model-calibration in consciousness science. If both arguments are sound, one is justified in holding a radically skeptical stance against the possibility of calibrating consciousness detection procedures. I now answer both arguments in turn.

3. An answer to the arbitrariness argument

The arbitrariness argument rests on a fallacy. The fallacy is to consider that one cannot calibrate a procedure by comparing its outcomes with those of a procedure that is known to be inaccurate. Here is, for instance, what Goldman (2002) writes:

> how are we supposed to identify any empirically observable variable that validates verbal reports? Whatever scientific variable, V, is found to correlate with verbal reports, this will not help validate the reports as a measure of consciousness *unless V itself is an accurate measure of consciousness*. But how can that be ascertained without relying on verbal reports? (Goldman, 2002, p.129, my emphasis).

That scientists can calibrate a procedure by systematically comparing its outcomes with those of a different, inaccurate procedure, is not obvious. But the claim that it can't be done should

sound suspicious. For one thing, if calibration always required accurate procedures in the first place, the whole exercise would be both impossible and futile. But it's not. So let me illustrate how concordance-calibration is achieved by borrowing a case from the history of science.

Scientists calibrated the first thermoscopes by comparing the outcomes of the thermoscopes with their own sensations of hot and cold (Chang, 2004, p.44-46). They knew perfectly well that sensation-based procedures for judging temperature are often inaccurate. But this didn't stop them. After all, they had no reason to believe that those judgments are *systematically inaccurate* in most situations. That is, researchers were *prima facie* warranted in believing the outcomes of sensation-based procedures in a large range of situations. And by showing that the outcomes of thermoscope-based procedures systematically agreed with the outcomes of sensation-based procedures in those ordinary cases, the former could inherit the small degree of epistemic standing of the latter.

But, you ask: 'how could scientists demonstrate that thermoscope-based procedures are *more* accurate than sensation-based procedures?' Good question. They did it by showing that thermoscope-based procedures gave more reliable results than sensation-based procedures in a range of cases in which they had reasons to believe that sensation-based procedures were unreliable.

'And how could they identify cases in which sensation-based procedures are unreliable without an independent standard for judging the accuracy of those procedures?' Short answer: they had pre-theoretical principles.

One principle is the principle of single value: something can't be in several incompatible states at the same time. Now if you put your right hand in a bucket of warm water, your left hand in a bucket of cold water, and then your two hands in a bucket of lukewarm water, you'll find that lukewarm water feels hot and cold at the same time. But water is not both hot and cold. So, there's something wrong with the sensation-based procedure in at least a given range of conditions. We know this because it provides outcomes that are incompatible with a principle that everyone – except perhaps toddlers and some quantum physicists – would regard as highly plausible.

Thermoscopes did not fall prey to the same problems (Chang, 2004). This means that thermoscope-based procedures were *as accurate* as sensation-based procedures over the range of cases in which researchers were *prima facie* warranted in believing the outcomes of

sensation-based procedures. And they were also *more* accurate than sensation-based procedures over the range of cases in which researchers had reasons to believe that sensation-based procedures were inaccurate. So, researchers determined the superior accuracy of thermoscope-based procedures by comparing their outcomes with those of a procedure that they knew to be inaccurate. Inaccurate procedures are not thereby good for nothing.

Essentially the same story goes in the case of consciousness detection procedures. The fact that researchers cannot use *d'* as an indication of consciousness in many cases does not imply that Type-1-based procedures cannot be used to calibrate other procedures.

Imagine the following task. You stare at a blank screen. Every five seconds or so, a big, clearly visible, picture of a squirrel appears at the center of the screen for three seconds. Your task is to press a button when you see a squirrel. In this very simple detection task, I see no good reason to believe that *d'* is not a good indication of consciousness of the pictures.

You might believe that this case is too trivial to be interesting. But please hold on. Here is another triviality: *d'* is a better indication of consciousness in this task than the number of times your eyes blinked during the task, your distance from the Eiffel Tower, or the rate of growth of your fingernails. Why is *d'* a better indication of consciousness than all those other possible indications? Why do Type-1-based procedures seem to have some *small degree of epistemic standing* that other possible procedures don't have?

Here's why. The better one sees something, the more likely one is to be conscious of it; and the worse one sees something, the less likely one is to be conscious of it[12]. I call this the "Basic Principle n°1". In this case, being 'better' or 'worse' at seeing something refers to the capacity of the perceptual system to extract some meaningful information about a stimulus that can be used by the subject. Given that *d'* can be interpreted as a measure of how well one sees stimuli in that sense, one is *prima facie* warranted in interpreting *d'* as an indication of consciousness, at least in a limited range of cases. The fact that *d'* conforms with Basic Principle n°1 gives some small degree of epistemic standing to Type-1-based procedures. This small

[12] If you're not convinced, try to do a discrimination task when stimuli are presented behind your head, and then do the same when stimuli are presented in front of you. I bet that *d'* will be higher in the latter case, and that, in that case, you will be conscious of the stimuli more often. If not, well done, you've just found a proof that you're gifted with extra-sensory perception, and that most of what we believe about perception is false. This puts you in a good position to become a superhero *and* win a Nobel Prize!

degree of epistemic standing is what separates Type-1-based procedures from any other procedure relying on arbitrary indications.

If the results of a given Type-2-based procedure do not follow the results of Type-1-based procedures at all, this constitutes a good reason to be suspicious of that procedure. For it would mean that one has found a case in which the Basic Principle n°1 is false. But it's likely to be true in most cases.

If the results of a Type-2-based procedure closely follow those of a Type-1-based procedure, the former inherits the small degree of epistemic standing of the latter. Agreement between those two types of procedures suggests that they are just as good at conforming with Basic Principle n°1. That's a first step toward concordance-calibration.

The upshot of this discussion is that there is nothing arbitrary in preferring procedures that follow the results of Type-1-based procedures to the procedures that do not. The former conform with a principle that seems *prima facie* plausible, while the latter don't.

But skeptics can reach deeper into their bag of tricks. Irvine (2019) writes:

> introspective evidence is only 'trusted' when it matches with non-introspective evidence, and ignored when it does not. Across all cases then, where introspective evidence either fits or does not fit with other evidence, introspective reports are not in a position to play a strong justificatory role in consciousness science, and are distinctly incapable of resolving long-standing theoretical and empirical debates. (p.20)

What's the point of using Type-2-based procedures if the only standard to evaluate them is to compare them to Type-1-based procedures? Scientists might as well directly use the gold standard Type-1-based procedures instead. Using Type-2-based procedures is either arbitrary or pointless.

The answer to this worry is that scientists are also warranted in believing that Type-1-based procedures are inaccurate *in some cases*. This is because Type-1-based procedures sometimes violate Basic Principle n°2: people can usually tell whether they are conscious of something or not. Once again, even if there are counterexamples to this principle in humans, they hopefully constitute exceptions rather than the rule. The Basic Principle n°2 has some small degree of epistemic standing.

As mentioned above, blindsight subjects exhibit above chance performance during discrimination tasks, but claim that they do not see anything in their blind field (Cowey, 2010;

Weiskrantz, 2009). Basic Principle n°2 gives a *prima facie* warrant to hold that blindsight subjects do not consciously see anything in their blind field. If they did, they would probably be able to tell. This also gives a good reason to believe that Type-1-based procedures are probably inaccurate *in this particular case*.

We now reach a crucial step. Researchers are *prima facie* warranted in interpreting *d'* as a relatively accurate indication of consciousness in a large range of cases. They are also warranted in believing that Type-1-based procedures are *not* accurate in some cases, like blindsight.

Now, if a Type-2-based procedure is as accurate as Type-1-based procedures when scientists are warranted in believing the results of Type-1-based procedures, but also conforms with Basic Principle n°2, then they are warranted in holding that this procedure is more accurate than Type-1-based procedures. Indeed, it is *as accurate* as Type-1-based procedures when they are warranted in believing that their results are accurate. And it is probably *more* accurate than Type-1-based procedures when they are *not* warranted in believing that their results are accurate. So, there is nothing arbitrary in using Type-2-based procedures instead of Type-1-based procedures to detect consciousness.

To be clear, I do not claim that concordance-calibration has already been successfully achieved in consciousness science. Instead, my claim is that there's nothing so special about *consciousness* as an object of scientific investigation when it comes to using concordance-calibration for calibrating detection procedures. Against the 'arbitrariness argument', I hold that concordance-calibration *can be* achieved for consciousness detection procedures, and that comparing various consciousness detection procedures is a meaningful and fruitful endeavour.

My argument above assumes that it is possible to find Type-2-based procedures that are as accurate as Type-1-based procedures over the range of cases in which scientists are warranted in believing the results of Type-1-based procedures. However, many sources of error could influence Type-2-based procedures, thus providing reasons to doubt the accuracy of their outcomes.

For this reason, the skeptics could grant that concordance-calibration is possible in principle, while also arguing that it is impossible *de facto*, given the many putative sources of errors affecting Type-2-based procedures.

I have two responses to this argument. First, I agree that concordance-calibration in consciousness science is hard. But calibrating detection and measurement procedures is almost always a dauntingly difficult task. The skeptics need to show that there is something special about *consciousness* that makes concordance-calibration of consciousness detection procedures *de facto* impossible. For otherwise they would have to declare that concordance-calibration is *de facto* impossible in all cases in which a large number of confounding factors could potentially affect detection and measurement procedures. Historical examples of successful concordance-calibration in such cases indicate that this claim is wrong[13].

Second, there are reasons for optimism. Let me illustrate this with an example. Empirical research indicates that when observers are not sure whether they saw a stimulus or not, as can happen for weak stimuli, they tend to answer that they do not see anything (Macmillan & Creelman, 2005). Type-2 reports, such as 'seen' or 'not seen', are contaminated by a 'conservative response bias'. Type-1-based procedures relying on the Signal Detection Theoretic indicator *d'* are protected against this source of error. This could be a valid reason to prefer Type-1-based procedures over Type-2-based procedures. However, in recent years this difficulty has been significantly reduced, as consciousness scientists have developed Type-2-based procedures that control for this confounding factor (Peters & Lau, 2015) or attempt to statistically reduce it (Maniscalco & Lau, 2012). While it is probably too early to confirm that we're done with this specific confounding factor, it is clear that calibration will only get better from here.

Of course, response biases are just one putative source of error among many others. A proper concordance-calibration would need to compare the outcomes of Type-1-based procedures to the outcomes of a variety of Type-2-based procedures that do, and do not, control for these confounding factors. Concordance-calibration can also consist in comparing various Type-2 procedures that are affected by different putative sources of errors, or by the same

[13] Take, for instance, Regnault and De Luc's long and tedious work of concordance-calibration of thermometers (Chang, 2004, p.76-84). Here's just a small, non-exhaustive list of the potential factors that could influence measurement outcomes in this case: Were air thermometers more accurate than mercury thermometers, or alcohol thermometers? How did the type of alcohol influence the outcome: were alcohol thermometers with old Languedoc wine more reliable than thermometers with Brandy, or mixtures of alcohol? How did the type of glass of the thermometer influence the outcomes: was Swedish glass better than green glass, or 'Choisy-le-Roi' crystal? Would air thermometers with different air densities provide different outcomes? What about thermometers with different gases, such as hydrogen, carbon dioxide or sulfuric acid gas? These questions were solved in large part through a long work of concordance-calibration involving systematic comparisons between different thermometers (Chang, 2004).

sources of errors to different degrees. By doing so, one can evaluate the extent to which putative sources of errors influence the outcomes, determine which procedure is more accurate, or develop alternative procedures.

This work of concordance-calibration is well underway in consciousness science, as several studies attempt to compare different report scales (e.g. Sandberg et al. 2010; Rausch et al., 2015, 2018; Szczepanowski et al., 2013). For example, concordance-calibration was used by Rausch & Zehetleitner (2014) to compare the outcomes of procedures using discrete, and continuous Type-2 report scales. This allowed them to rule out a possible influence of the nature of the scale on Type-2 reports.

In sum, comparing various detection procedures influenced by different sources of errors, or by the same sources of errors to different degrees, is surely a long and difficult task. But that's not very surprising: calibrating detection and measurement procedures is difficult. Yet, historical examples indicate that successful concordance-calibration has been achieved even in cases where many potential confounding factors could influence measurement procedures. In addition, there is evidence that scientists have already made some progress in calibrating Type-2-based procedures in order to diminish the influence of confounding factors. Against the skeptics, I conclude that concordance-calibration in consciousness science is not only possible in principle, but also possible *de facto*, and actually underway.

4. An answer to the crazy spaghetti argument

Remember, the crazy-spaghetti argument goes like this: model-calibration requires a model of the detection procedure itself. But introspection, which plays a key role in Type-2-based procedures, is too complex to be modeled. Hence, scientists cannot calibrate Type-2-based procedures. In what follows I explain why this argument is unsound. I conclude with a case study of a successful model-calibration.

4.1. Random and systematic detection errors

The good news is that scientists don't have to model all the factors that could potentially influence introspection. They just have to model the factors that would lead them to commit *systematic* detection errors.

Following the distinction between systematic and random measurement errors in the *International Vocabulary of Metrology* (Joint Committee for Guides in Metrology, 2012), we can identify two main sources of detection errors. I define a systematic detection error as a source of detection error that in replicate detections remains constant or varies in a predictable manner. A random detection error is a source of detection error that in replicate detections varies in an unpredictable manner.

An example of systematic error first. Imagine a Type-2 task in which subjects provide reports of the kind "I saw the stimulus"; or "I did not see the stimulus". If a subject has a conservative bias, that is, if the subject tends to use the "unseen" report category regardless of her actual conscious experience, scientists could have the tendency to interpret her reports as indicating that she was almost never conscious of the stimuli. On repeated detections the subject's conservative bias is likely to remain the same, and scientists will thereby commit the same detection error. The subject's bias constitutes a source of systematic detection errors.

Now for a case of random error. Over the course of hundreds of trials, subjects could sometimes press response buttons that they didn't mean to press. These kinds of events could also lead to detection errors. But there's no reason to think that this error will persist on repeated detections. Hopefully, subjects do not always make these kinds of mistakes. And even if they often did, there is no reason to believe that these errors will drive the experimenter's interpretation of the indications in any determinate direction.

Random sources of errors are impossible to completely eliminate. Not only is it difficult to determine whether sources of random errors influence the results, but it is also difficult to determine *how* they influence the results. Since their effects on the detection procedure outcomes can hardly be modelled, random errors are particularly problematic for model-calibration.

Fortunately, it is quite easy to reduce random errors. Since they are random, repeating the detection process a large number of times should cancel them. That's precisely what consciousness scientists do. In Section 1, I insisted on the fact that consciousness scientists do not detect consciousness on a trial-by-trial basis, but infer consciousness *across* many trials, based on a *pattern* of Type-1 and Type-2 reports. In the long run, this method largely cancels the effects of random sources of errors.

Another important point is that some sources of errors that are systematic at the level of a single subject can be randomly distributed *across* subjects. To understand, take the following

analogy. I want to measure the length of a table. To do so I have a set of different rulers. Each ruler might have its own source of systematic errors. To avoid my result being "contaminated" by the idiosyncratic systematic errors of a given ruler, I can measure the table with each ruler, and then average the results. Provided that the systematic sources of errors are randomly distributed across the different rulers, the final measurement outcome should be closer to the true value of the length of the table.

In the same way, response biases, for instance, could cause systematic detection errors. But in the *group* of subjects participating in the experiment, participants are likely to have different biases. If the results are aggregated, and assuming that biases are more or less randomly distributed across subjects, the systematic errors present at the level of each individual should cancel each other out.

In consciousness science, detection procedures typically output a judgment about consciousness of the stimuli, *across trials*, for a *group* of subjects. By doing so, scientists can avoid the contamination of their detection procedures by individual differences in the ways in which the subjects react to the experimental setting. These kinds of systematic errors, randomly distributed among subjects, should thereby tend to be erased by using the detection procedure on groups of subjects, instead of individual subjects.

Nevertheless, if using the procedure on many subjects performing many trials can diminish the effects of random errors, all confounding factors cannot be reduced in this way. Some systematic errors will remain, even in those cases.

For instance, we just assumed that response biases could be randomly distributed across subjects. But in detection tasks participants often implicitly aim at minimizing the number of false alarms to a predetermined level, a goal known as the Neyman-Pearson objective (Green and Swets, 1966). When sensitivity is low, adopting the Neyman-Pearson objective leads to a conservative bias: to minimize false alarm rates, participants become more conservative as sensitivity diminishes. If subjects adopt the Neyman-Pearson objective, biases should remain a systematic source of error even when random errors have been reduced by judging the subjects' consciousness across trials, and for groups of subjects[14].

[14] In practice, there are two main ways in which consciousness scientists attempt to mitigate the effects of response biases. The first is to use statistical tools inspired from Signal Detection Theory such as meta-d' (Maniscalco & Lau, 2012), or the area under the Type-2 ROC curve (AUROC2) (Fleming & Lau, 2014).

The good news is that, while the influence of random errors on the detection outcome cannot be modelled, the influence of *systematic* errors can. That's where scientists appeal to model-calibration. Let me now show how this is done with a case study.

4.2. Model-calibration in consciousness science: A case study

An experiment by Fleming and Dolan (2010) is a good example of model-calibration in consciousness science. Before presenting it, I have to give a quick introduction to post-decision wagering procedures.

In post-decision wagering, subjects perform the Type 1 task and then place a bet on whether their response on the Type 1 Task was correct (Figure 1; Koch and Preuschoff, 2007; Persaud et al., 2007). If the participant saw the stimulus very well and is confident that she did the task correctly, she should be willing to bet on it. If she didn't see the stimulus consciously, she'll probably go with the low bet instead. So, the betting behavior indicates something about the subject's consciousness of the stimuli.

Why would anyone want to assess consciousness in this way? Because it solves three problems that other detection procedures might have. First, the incentive problem. Subjects don't have much incentive to provide accurate confidence or visibility judgments. That's not a problem with post-decision wagering: people like to earn money. Second, children and non-human animals don't have a clue about how to use confidence and visibility ratings. They nevertheless understand under which conditions they can earn some rewards[15]. Finally, subjects can interpret confidence and visibility categories in many different ways. Post-decision wagering response categories are easy to understand unequivocally.

The second approach is to design "bias-free" tasks, such as two-interval forced choice tasks (de Gardelle & Mamassian, 2014; Knotts et al., 2018; Peters & Lau, 2015; Peters et al., 2017).

[15] For instance, in a study by Stolyarova et al. (2019) rats were trained to discriminate the orientation of Gabor patches. After their decisions, they could either directly initiate a new trial, or wait for a sugar pellet reward, provided only if their decision was correct. The basic idea is that if rats are unsure about their decisions they should directly initiate a new trial instead of wasting time waiting for a sugar pellet that they probably won't get.

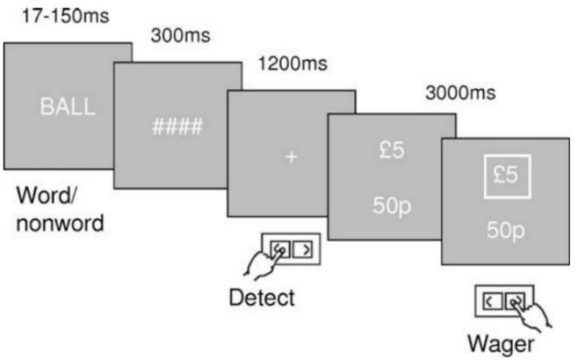

Figure 1. Source: Fleming and Dolan (2010). Post-decision wagering procedure. The subject performs the Type 1 task (here, identifying whether a series of letters is a word or not). She then has to bet either 5 pounds or 50 pennies on the correctness of her decision. Depending on the accuracy of her decision, she either earns or loses 5 pounds or 50 pennies.

But there's no such thing as a methodological free lunch. The downside of post-decision wagering is the threat of loss aversion (Kahneman, 2011). The pain of losing is greater than the satisfaction of winning: when placing bets people typically attempt to avoid losses and optimize their bets for sure wins. If subjects are not confident enough but still somewhat confident, they might prefer to choose the low bet instead of the high bet. Post-decision wagering could lead subjects to adopt a conservative bias.

Perhaps that's true. Perhaps not[16]. If not, scientists would miss on a promising consciousness detection procedure. They need to evaluate the detection errors that this procedure could lead them to make in order to avoid rejecting it on the basis of the mere suspicion that the method is affected by loss aversion—they need to calibrate it.

That's what Fleming & Dolan (2010) did by using *ideal observer models*. In the Type-1 task of this experiment, subjects had to decide whether a stimulus quickly followed by a mask was a word or not (Figure 1). The Type-2 task was a post-decision wagering task, in which subjects had to bet on the accuracy of their Type-1 decisions.

Signal Detection Theory models the Type-1 decision as being based on two overlapping Gaussian distributions – the noise (non-word), and the signal (word) distributions – over a

[16] Recent literature in social and consumer psychology casts doubt on the existence of loss aversion as a systematic bias (for a review, see Gal and Rucker, 2017). In general, the effect of loss aversion might be smaller than usually assumed (Walasek et al., 2018). Instead, it could be that the low bet is perceived by subjects as maintaining the status quo, and subjects have a status quo bias instead of loss aversion (Gal, 2006).

stimulus axis, often interpreted as reflecting 'signal strength'. On each trial, the Type-1 decision ultimately depends on the strength of the sensory signal, relative to the position of the participant's response criterion (Figure 2a). Since the signal and noise distributions overlap, subjects aren't always right. Words are sometimes misclassified as non-words, and vice versa.

Given signal and noise distributions, scientists can determine the probability of responding correctly or incorrectly on a given trial (Galvin et al. 2003). For instance, the probability of being incorrect is higher if signal strength falls where the two distributions overlap (Figure 2b). Thus, for any given Type-1 performance, scientists can model *ideal* Type-2 behaviors that maximize wagering performance. That is, for any Type-1 performance, one can model the responses that one should expect if subjects are unaffected by biases (Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012; Peters & Lau, 2015).

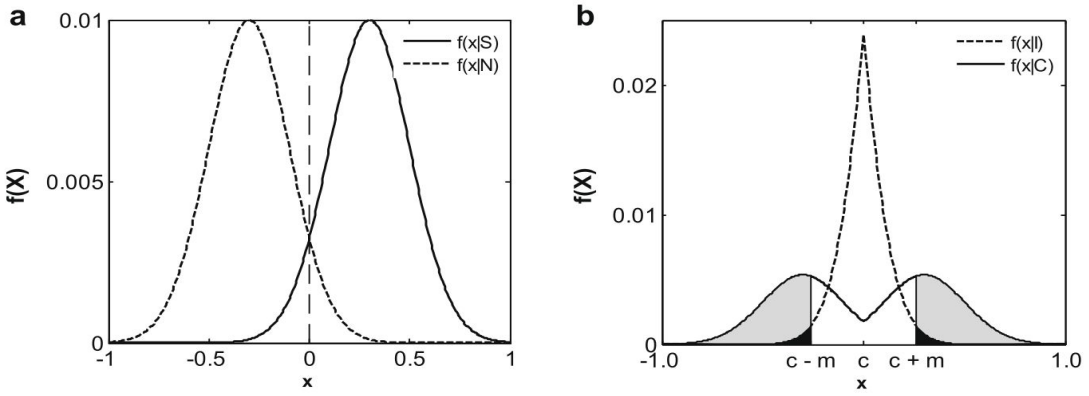

Figure 2. (a) Dashed line: noise distribution. Solid line: signal distribution. The distributions are drawn over a random variable x corresponding to the stimulus axis. The dashed line between the two distributions is the optimal response criterion. (b) Probability density for making a correct (solid line) or incorrect (dashed line) response for different values of x.

But optimality is a relative notion: one has to define *with respect to what* a given behavior is optimal. Here, it was assumed that observers attempted to maximize utility, or the value gained as a result of each trial. Each outcome on the Type-2 task was associated with a given utility gain or loss, resulting in a utility function specifying how much utility a given monetary reward, or monetary loss, was worth to the observer. The ideal observer was modelled with a symmetric utility function: monetary gains and losses were attributed the same weight. Loss averse observers, on the other hand, were modelled with asymmetric utility functions: monetary losses were attributed a greater weight than monetary rewards.

Based on these models, Fleming & Dolan could thus simulate the Type-2 behavior of an ideal observer, and that of a loss averse observer, for any given Type-1 performance. The models predicted that loss aversion would reduce the observers' propensity to opt for the high wager independently of the correctness of their responses on the Type 1 task. This tendency would increase when a lot of money is at stake and decrease otherwise.

The next step was to test the predictions of the models against actual psychophysical data. Subjects performed a Type 1 discrimination task and a Type 2 task involving post-decision wagers. Fleming & Dolan manipulated the subjects' loss aversion by presenting them with wagers of different sizes.

The results indicated that the performances of participants closely followed those predicted by the loss-averse observer model. But there's more. Fleming & Dolan could use their model to estimate the loss aversion parameter for each subject to determine exactly how individual performances on the Type 2 task were affected by loss aversion.

That's a successful evaluation of the detection procedure. The demonstration that post-decision wagering is affected by loss-aversion led to the subsequent development of "no-loss" post-decision wagering procedures with the goal of mitigating the effects of the loss-aversion factor (Dienes & Seth 2010; Mealor & Dienes, 2012; see also Clifford et al. 2008).

This study illustrates how experimenters can evaluate which factors influence detection procedures by using models predicting the subjects' performances *if they were influenced by those factors*, compared to the performance that they would have if they were ideal observers. In most cases, model-calibration in consciousness science consists in building ideal observer models, and adding parameters to the models reflecting the different factors that might influence the subjects' Type-2 performances, given their Type-1 performance.

4.3. More reasons for optimism

But scientists can't model sources of errors that they don't even suspect to influence detection outcomes, or sources of errors that influence detection outcomes in yet unknown ways. The more 'crazy-spaghetti' Type-2 decision-making is, the harder model-calibration is going to be. And the harder model-calibration is, the stronger the case for skepticism.

Type-2 decision-making is indeed suboptimal in many ways (for a review, see Rahnev & Denison, 2018). Nevertheless, there are two main reasons to remain optimistic about the progress of calibration in consciousness science.

First, sources of detection errors are increasingly well understood. At least in the case of confidence judgments, scientists have successfully identified many potential sources of errors (Fleming & Daw, 2017; Rahnev & Denison, 2018).

Identifying potential sources of errors allows researchers to evaluate which confounding factors should be controlled for, and which sources of errors are *de facto* statistically reduced. For instance, Rahnev & Fleming (2019) have recently suggested that 'staircasing procedures', which consist in adjusting stimulus intensity to maintain a given level of performance throughout an experiment, could artificially inflate metacognitive sensitivity, thus leading to systematic detection errors. As a result of this discovery, scientists can improve their procedures by using constant-stimulus designs, or analyzing performance on each stimulus intensity level separately. As this example illustrates, the study of Type-2 decision-making and the improvement of detection procedures will go hand-in-hand. The good news is that Type-2 decision-making is now extensively studied, and increasingly well understood (See, e.g., Fleming & Daw, 2017).

Second, and perhaps most importantly, perfect accuracy is a perfectly unreasonable standard. A procedure should only aim to be accurate *enough* given the researchers' epistemic and practical goals (Elgin, 2017). In a nutshell, a detection procedure is accurate *enough* for a given epistemic or practical goal if a more accurate procedure would have provided the same detection outcome. In other words, a detection outcome is accurate enough when its inaccuracy does not undermine its epistemic and practical functions.

Let me illustrate this with an example. Rounis et al. (2010) aimed to analyze the effect of continuous theta-burst stimulation (cTBS) to the prefrontal cortex on consciousness. They did so by comparing behaviors from subjects before and after cTBS, as well as subjects who received real versus sham cTBS. They observed that, on average, consciousness of stimuli decreased for subjects who received real cTBS, compared to before cTBS, and after sham cTBS.

In this case, all sources of detection errors do not necessarily need to be controlled for. Instead, all that is required is that sources of detection errors present in one group of subjects are equally present in the other group of participants, as well as in both conditions of the experiment. Assuming that those sources of detection errors were indeed similar in the different

conditions of the experiment, the detection outcome wouldn't have been different even if Rounis et al. had used a more accurate procedure. Under this assumption, the procedure was accurate *enough* for their purposes.

Since detection procedures only have to be accurate enough in that sense, the mere existence of putative sources of detection errors is not sufficient to justify radical skepticism. One also needs to demonstrate that a different detection outcome would have been obtained if sources of detection errors had been controlled for. This, it seems, can only be decided on a case-by-case basis.

To conclude, the skeptical worries of the crazy spaghetti argument seem unwarranted. First, a wide variety of errors are already reduced through statistical analyses, since there is no reason to hold that those errors should influence the detection outcomes in *systematic* ways. Detection outcomes are obtained based on statistical analyses of the responses of groups of subjects collected over multiple trials. Second, the remaining systematic errors can be modelled through ideal observer models, thereby allowing model-calibration. Subsequent studies can then attempt to mitigate the influence of the factors that have been identified as sources of errors through model-calibration. Finally, I provided reasons for optimism about the calibration of consciousness detection procedures, *even if* Type-2 decision-making is indeed a cognitive confluence of 'crazy-spaghetti'.

Conclusion

There is still a lot of work to do to have accurate and valid consciousness detection procedures. My goal here was merely to argue, against the skeptics, that this work *can* be done, and is worth doing. Consciousness detection procedures can be calibrated. Future research should attempt to determine the conditions under which those procedures are accurate, explain why they sometimes provide different outcomes, and assess what those procedures detect exactly (Michel, 2019b; Rosenthal, 2019). I suggested that this work could be carried out both through systematic comparisons of the outcomes of different procedures, as well as by developing more accurate models of the procedures themselves.

Acknowledgments: I thank Jorge Morales, Hakwan Lau, Megan Peters, Liz Irvine, Steve Fleming, Keith Frankish, as well as two anonymous reviewers for their comments on this paper.

References

Basso, A. (2017). The appeal to robustness in measurement practice. *Studies in History and Philosophy of Science Part A*, *65–66*, 57–66.

Bayne, T., & Spener, M. (2010). Introspective Humility. *Philosophical Issues*, *20*, 1–22.

Carruthers, P., & Ritchie, J. B. (2012). *The emergence of metacognition: Affect and uncertainty in animals*. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition* (p. 76–93). Oxford University Press.

Chang, H. (2004) *Inventing Temperature: Measurement and scientific progress*. Oxford University Press.

Cheesman, J., & Merikle, P. M. (1986). Distinguishing Conscious from Unconscious Perceptual Processes. *Canadian Journal of Psychology*, *40*(4), 343–367.

Chirimuuta, M. (2014). Psychophysical Methods and the Evasion of Introspection. *Philosophy of Science*, *81*(5), 914–926.

Clifford, C. W. G., Arabzadeh, E., & Harris, J. A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences*, *12*(2), 54–58.

Cowey, A. (2010). The blindsight saga. *Experimental Brain Research*, *200*(1), 3–24.

de Gardelle, V., Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, 25,1286–1288.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1322–1338.

Dienes, Z., & Perner, J. (2004). Assumptions of a subjective measure of consciousness: Three mappings. In R. Gennaro (Ed.). *Higher order theories of consciousness* (pp. 173–199). Amsterdam: John Benjamins Publishers.

Dienes, Z. and Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task.

Elgin, C. (2017). *True Enough*. Cambridge, MA: MIT Press.

Feest, U. (2014). Phenomenal Experiences, First-Person Methods, and the Artificiality of Experimental Data. *Philosophy of Science*, *81*(5), 927–939.

Feest, U. (2012). Introspection as a method and introspection as a feature of consciousness. *Inquiry*, *55*(1), 1–16.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

Fleming, S. M. and Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and Cognition*, 19(1):352–363.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(July), 443.

Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, 1(1):23–32.

Gal, D. and Rucker, D. D. (2017). The Loss of Loss Aversion: Will It Loom Larger Than Its Gain? *Journal of Consumer Psychology*, 28(3), 497-516.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability : Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

Goldman, A. (2002). *Pathways to Knowledge: Private and Public.* Oxford University Press.

Goldman, A. I. (2004). Epistemology and the evidential status of introspective reports. *Journal of Consciousness Studies*, *11*(7–8), 1–16.

Green, D. and Swets, S. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hacking, I. (1981). Do we see through a microscope? *Pacific Philosophical Quarterly*, 62(4):305– 322.

Holender, D. (1986). Semantic activation without conscious identification in dichotic-listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences*, 9, 1–23.

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

Irvine, E. (2012b). *Consciousness as a Scientific Concept*. Springer.

Irvine, E. (2013). Measures of consciousness. *Philosophy Compass*, *8*(3), 285–297.

Irvine, E. (2019). Developing Dark Pessimism Towards the Justificatory Role of Introspective Reports. *Erkenntnis*, (0123456789).

Joint Committee for Guides in Metrology (JCGM). (2012). *International Vocabulary of Metrology (VIM), 3rd edition.* http://www.bipm.org/en/publications/guides/vim.html.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books.

Knotts, J. D., Lau, H., & Peters, M. A. K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. *Attention, Perception, & Psychophysics*, *80*(8), 1974–1987.

Kroker, K. (2003). The progress of introspection in America, 1896–1938. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *34*(1), 77–108.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2012). Robustness analysis disclaimer: Please read the manual before use! *Biology and Philosophy,* 27, 891-902.

Lau, H. (2008). Are we studying consciousness yet? In M. Davis & L. Weiskrantz (Eds.), *Frontiers of Consciousness* (pp. 1–22).

LeDoux, J. (2019). *The Deep History of Ourselves: The Four-Billion-Year Story of How We Got Conscious Brains.* Viking.

Li, Q., Hill, Z., & He, B. J. (2014). Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. *The Journal of neuroscience*, *34*(12), 4382–4395.

Lyons, W. (1986). *The disappearance of introspection*. Cambridge, MA: MIT Press.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Taylor & Francis.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data : Meta- d', Response- Specific Meta-d', and the Unequal Variance SDT Model. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 25–66).

Mealor A, Dienes Z (2012) No-loss gambling shows the speed of the unconscious. *Consciousness and Cognition*, 21(1):228–237.

Michel, M. (2019a). Consciousness Science Underdetermined: A short history of endless debates. *Ergo*.

Michel, M. (2019b). The Mismeasure of Consciousness: A problem of coordination for the Perceptual Awareness Scale. *Philosophy of Science*, 86(5), 1239-1249.

Michell, J. (1999). *Measurement in Psychology: A critical history of a methodological concept*. Cambridge University Press.

Morales, J., Chiang, J., & Lau, H. C. (2015). Controlling for performance capacity confounds in neuroimaging studies of conscious awareness. *Neuroscience of Consciousness*, *2015*(1).

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

Norman, E. and Price, M. C. (2015). Measuring consciousness with confidence ratings. In Overgaard, M., editor, *Behavioral Method in Consciousness Research*. Oxford: Oxford University Press.

Overgaard, M., Rote, J., Mouridsen, K., and Ramsoy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and Cognition*, 15(4):700–708.

Persaud, N., McLeod, P., and Cowey, A. (2007). Post-decision wagering objectively measures awareness. Nature neuroscience, 10(2):257–61.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *ELife*, *4*(October), 1–30.

Peters, M. A. K., Fesi, J., Amendi, N., Knotts, J. D., Lau, H., & Ro, T. (2017). Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex*, *97*, 119–132.

Piccinini, G. (2009). First-Person Data, Publicity, and Self-Measurement. *Philosophers' Imprint*, 9(9).

Rahnev, D., & Denison, R. N. (2018). Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, *41*, 1–107.

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *5*(April), 1–42.

Rausch, M. and Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition*, 28:126–140.

Rausch, M., Müller, H. J., and Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition*, 35:192– 205.

Rausch, M., Hellmann, S., and Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, perception & psychophysics*, 80(1):134–154.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175.

Rosenthal, D. (2018). Consciousness and confidence. *Neuropsychologia*, 128, 255-265.

Sandberg, K. and Overgaard, M. (2015). Using the perceptual awareness scale (PAS). In Overgaard, M., editor, *Behavioral Methods in Consciousness Research*. Oxford: Oxford University Press.

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.

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

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728.

Schwitzgebel, E. (2011). Perplexities of consciousness. Cambridge: MIT Press.

Schwitzgebel, E. (2012). Introspection, what? In D. Smithies & D. Stoljar (Eds.), Introspection and consciousness (pp. 29–48). Oxford: Oxford University Press.

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception: A model-based approach to method and evidence. *Perception and Psychophysics*, 66(5), 846–867.

Spener, M. (2013). Moderate scepticism about introspection. Philosophical Studies, 165(3), 1187.

Spener, M. (2015). Calibrating introspection. Philosophical Issues, 25(1), 300–321.

Spener, M. (forthcoming). Consciousness, introspection, and subjective measures. In U. Kriegel (Ed.), *Handbook of the Philosophy of Consciousness*. Oxford University Press.

Stolyarova, A., Rakhshan, M., Peters, M.A.K., Lau, H., Soltani, A., & Izquierdo, A. (2019). Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nature Communications, 10*(4704).

Szczepanowski, R., Traczyk, J., Wierzchoń, M., and Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. *Consciousness and Cognition*, 22(1):212– 220.

Tal, E. (2014). Making time: A study in the epistemology of measurement. *British Journal for the Philosophy of Science*, 67(1):297–335.

Tal, E. (2017). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A*, *65–66*, 33–45.

Timmermans, B. and Cleeremans, A. (2015). How can we measure awareness? an overview of current methods, in M. Overgaard (ed.), *Behavioral Methods in Consciousness Research* (Oxford University Press).

Titchener, E. B. (1905). *Experimental psychology: A manual of laboratory practice*. New York: The Macmillan Company.

Walasek, L., Mullett, T. L., & Stewart, N. (2018). A meta-analysis of loss aversion in risky contexts. Available at SSRN: https://ssrn.com/abstract=3189088.

Weiskrantz, L. (1998), Consciousness and Commentaries. *International Journal of Psychology*, 33: 227-233.

Weiskrantz, L. (2009). *Blindsight: A case study spanning 35 years and new developments*. Oxford: Oxford University Press.

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., and Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, 27(1):109–120.

Zehetleitner, M. and Rausch, M. (2013). Being confident without seeing: What subjective measures of visual consciousness are about. *Attention, Perception, and Psychophysics*, 75(7):1406–1426.