PhiMiSci
Philosophy and the Mind Sciences

Distinguishing absence of awareness from awareness of absence

Matan Mazor[a] (mtnmzor@gmail.com)
Stephen M. Fleming[a,b] (stephen.fleming@ucl.ac.uk)

Abstract
Contrasting brain states when subjects are aware compared to unaware of a presented stimulus has allowed researchers to isolate candidate neural correlates of consciousness. Here we propose that an important next step in this research program is to investigate, perhaps paradoxically, brain states that covary with reports of absences of awareness. Specifically, we propose that in order to distinguish content-specific and content-invariant neural correlates of consciousness, a distinction needs to be made between the neural correlates of awareness of stimulus absence, and the neural correlates of absence of awareness (of either stimulus presence or absence). We ground this distinction in higher-order computational models of consciousness, where the state of higher-order nodes is invariant to the specific contents of awareness. To map the different levels of these models to neurophysiological correlates, we suggest two empirical approaches – inverted designs and two-dimensional awareness reports – in which reports about awareness and stimulus presence can be dissociated.

Keywords
Absence · Awareness · Consciousness · NCC

This article is part of a special issue on "The Neural Correlates of Consciousness", edited by Sascha Benjamin Fink.

1 Introduction

Contrasting brain states in trials where subjects are aware or unaware of a physical stimulus (the Aware-Unaware contrast; A-U; also known as *Contrastive Analysis*; Baars, 1993) has been a crucial tool for the study of Neural Correlates of Consciousness (NCC). In a typical experiment, participants' brain activity is monitored while

[a]Wellcome Centre for Human Neuroimaging, UCL

[b]Max Planck UCL Centre for Computational Psychiatry and Ageing Research; Department of Experimental Psychology, UCL

Mazor, M., & Fleming, S. M. (2020). Distinguishing absence of awareness from awareness of absence. *Philosophy and the Mind Sciences*, *1*(II), 6. https://doi.org/10.33735/phimisci.2020.II.69

they observe a series of stimuli presented at or near perceptual threshold. Contrasting average neural activation during trials in which subjects did or did not have conscious awareness of the stimulus is then taken to capture the neural processes that contribute to perceptual awareness above and beyond what is necessary for subliminal processing. In the past two decades the introduction of clever experimental manipulations such as meta-contrast masking (Lau & Passingham, 2006) and no-report paradigms (Pitts et al., 2012) has afforded further isolation of the effects of differences in conscious experience from the neural correlates of task relevance, task performance, and explicit report.

Inspired by the A-U contrast, computational models of visual awareness have in turn provided theoretical frameworks for the interpretation of empirical findings, focusing on what makes some stimuli available to consciousness while others remain unconscious. Baars (1993) proposed that conscious awareness is the result of a broadcasting of information to a global neural workspace. Drawing an analogy between consciousness and a working theatre, at any given moment only some of the actors (percepts, thoughts, memories, etc.) on a stage (working memory) are under the spotlight (attention) and visible to the audience (other cognitive faculties). This descriptive cognitive model was later grounded in candidate neural implementations, constrained by neurophysiological data (Aru et al., 2019; Dehaene et al., 2003). Over the years, these models have been refined and challenged by findings of increased activity in a widespread frontoparietal network in studies of the A-U contrast on the one hand (consistent with implementation of long-range neural connections supporting the global workspace; Del Cul et al., 2007; Dehaene & Changeux, 2011), and through debates over the extent of changes in consciousness in patients with prefrontal lesions on the other hand (see Koch et al., 2016; Boly et al., 2017; Odegaard et al., 2017, for recent reviews).

In parallel to these empirical and theoretical advances, researchers and philosophers have identified distinct subdivisions within the overarching terminology of the NCC. Marvan and Polak, in this volume, provide a comprehensive overview of recent NCC classification schemes, highlighting the important distinction between content-specific and general NCCs. In this article, we focus on a related distinction between content-specific and content-invariant NCCs (also resembling the distinction made between differentiating and non-differentiating NCCs; Bayne & Hohwy, 2013). We use these terms to refer to NCCs that are specific or invariant to the intentional contents of consciousness. Examples of content-specific NCCs are the neural correlates of being aware of a face, or the neural correlates of being aware of a stimulus on a computer screen. They are content-specific because they tell us something about the contents of experience (the presence of a face in the first case, or of a stimulus in the second). On the other hand, content-invariant NCCs include neural markers that distinguish different global or local states of awareness, independent of the intentional content of experience. These content-invariant states of awareness may fluctuate even within experimental paradigms that are designed to measure the content of consciousness, and with fully awake and sober healthy participants. For instance, changes in neural excitability, attention, or beliefs about attention can affect reports of stimulus awareness, even if such factors are independent of how a stimulus is represented or encoded. As we will see later, content-invariant NCCs can in some cases carry content (such as subpersonal beliefs about one's attentional state), as long as this content is not the content of consciousness.

As pointed out by others (Aru et al., 2012), the A-U contrast used in experimental research does not exclusively identify either content-specific or content-invariant aspects of the NCC. The fact that brain region X is more activated in trials where participants were aware of the stimulus may reflect a difference in conscious perceptual contents (stimulus presence or absence), and/or a difference in other content-invariant properties (for example, level of receptiveness to incoming sensory information, and/or meta-level beliefs about sensory processing, which may fluctuate and change even when awake and performing a task).

The distinction between these two interpretations is especially clear when zeroing in on trials in which subjects report being unaware. Let's take a closer look at what the participant might be communicating when they report being unaware of the target. We will put aside cases in which participants' responses do not reflect their actual state of awareness, for example due to motor errors in response. In the remaining veridical 'unaware' trials, participants may be expressing something along the lines of "Even though I was aware of other objects and events (for example the background), the target was not in the content of my awareness" (content-specific interpretation; *awareness of absence*), or alternatively something more like "My awareness level was low" (content-invariant interpretation; *absence of awareness*). This content-invariant interpretation can be further subdivided into lapses of awareness that are global ("My global awareness level was low") and local ("My awareness of visual stimuli in the fovea of my visual field was low").[1] Borrowing from Baar's theatre analogy, the participant may be reporting that there was no actor under the spotlight (content-specific), that the spotlight was dimmed or turned off (content-invariant, global), or alternatively that the spotlight was on but directed away from where the actor was supposed to stand (content-invariant, local).

In the remainder of this article, we will argue that in order to better understand the commonalities, differences, and interactions between content-specific and content-invariant NCCs, more research is needed into the different causes of 'unaware' responses. In what follows we will unpack what this may mean both theoretically and practically. Theoretically, in Section 2 we describe first- and higher-order approaches to the modeling of awareness reports, and explain how these approaches handle the difference between awareness of absence and absence of awareness. Practically, in Section 3 we propose two experimental approaches to correlate neural activity with distinct components of higher-order computational models: inverted designs and two-dimensional reports.

[1] We thank an anonymous reviewer for pointing out this important distinction.

2 Awareness of absence is not absence of awareness

A powerful approach to the modeling of awareness reports is grounded in decision theory, and specifically in Signal Detection Theory (SDT; Green & Swets, 1966). In its simplest form, SDT treats percepts as scalar values on a one-dimensional 'evidence' axis. Evidence in SDT can be thought of as mean-field activation in some sensory brain region or as the firing rate of single neurons, although no commitments are made to specific implementations. Evidence tends to be higher when a stimulus is present and lower when a stimulus is absent, but only probabilistically so: sometimes a stimulus leads to weak evidence, or the absence of a stimulus to strong evidence. Subjects then generate awareness reports by comparing the sensory evidence sample to an internal criterion. Only samples that fall above the criterion are reported as reflecting awareness to stimulus presence (see Fig. 1, panel a).

Already in its basic form, SDT successfully accounts for several core observations. First, not all percepts reach awareness, and sometimes illusory awareness of a stimulus occurs without a stimulus being present. These two error types are known as 'misses' and 'false alarms', respectively. Second, the ability to differentiate between the presence and absence of a stimulus and the overall tendency to report being aware of a stimulus may vary independently. The former is modeled as the shape of and distance between the distributions of strength of evidence when a stimulus is present or absent (sensitivity, or d'), and the second as the position of the criterion with respect to these distributions (c), both of which can be straightforwardly estimated from empirical data if equal-variance Gaussians are assumed.

This simple SDT model has been further extended to account for the relation between subjective reports and objective discrimination. For example, multi-dimensional SDT models treat percepts as points on a multi-dimensional space, with different dimensions representing different stimulus properties such as color or size (see Fig. 1, panel b). King and Dehaene (2014) have shown that such a model accounts for central findings on visual awareness, such as above-chance discrimination of stimuli that are reported as unseen, and the nonlinear relation between subjective visibility reports and sensory strength. Importantly, the multi-dimensional SDT model accounted for these findings without postulating additional processes supporting conscious awareness.

In the tradition of 'perception as inference' (Helmholtz, 1948), the application of SDT models in consciousness research implicitly equates perceptual awareness reports with subpersonal beliefs about the presence of signal. Similarly, reports of being unaware of a stimulus are modeled as reflecting a high posterior probability of a stimulus being absent. Importantly, these models do not differentiate between having no experience of a signal, and having an experience of no signal. Both are modeled as the result of a probabilistic inference where the winning hypothesis is

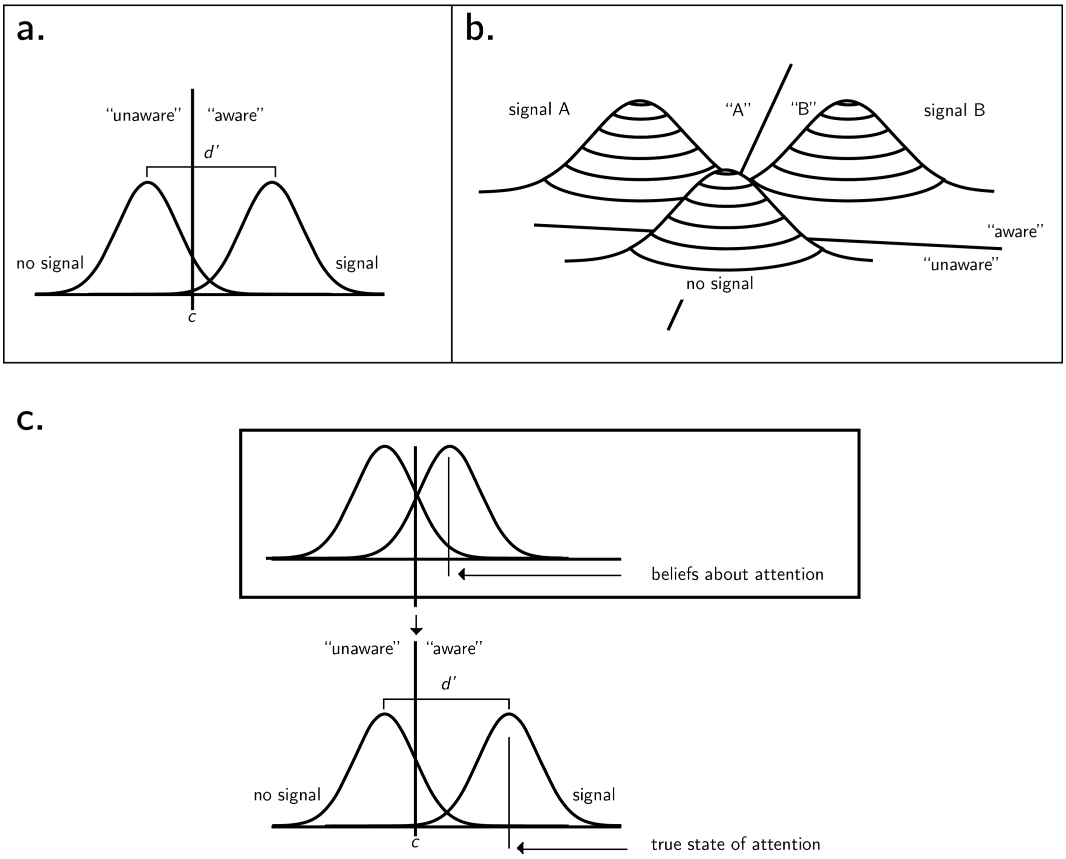

Figure 1: SDT models. of awareness reports. (a) In its traditional one-dimensional form, SDT describes awareness reports as a comparison of sensory samples against a decision criterion. Sensory samples probabilistically vary as a function of signal presence. (b) In multi-dimensional SDT, samples vary along more than one dimension, and different decisions can be made by applying different decision criteria. (c) In higher-order models, awareness reports are affected by the sensory samples and by beliefs about the underlying distributions. For example, the decision criterion shifts as a function of one's beliefs about their current attention state.

'stimulus absence'. For a computational model to make the important distinction between absence of awareness and awareness of absence that we identify above, it must allow for content-invariant and content-specific aspects of awareness to independently vary. Only then can a distinction be made between trials in which the participant was not aware of the target nor of its absence (low content-invariant consciousness), and trials in which they were clearly aware of the absence of the stimulus (high content-invariant consciousness in the absence of content-specific consciousness of the stimulus). In what follows, we provide examples of multi-dimensional or multi-layered computational architectures, and explain how they handle the distinction between absence of awareness and awareness of absence.

One example of such a multi-dimensional computational architecture is provided by the *Attention Schema Theory (AST)* of consciousness (Graziano, 2013;

Graziano & Webb, 2015). In AST, the proposition "I am aware of *X*" can be broken down into three parts: knowledge of myself (denoted *S*), knowledge of *X*, and finally, a schematic knowledge of the relation of 'being aware of' (denoted *A*). While states of awareness are critically dependent on *A* (and more specifically, on the schematic description of one's own mental attributes, including attention), in the AST framework they are not dependent on any specific *X* (Graziano, 2013, ch. 9). In other words, awareness of absence may be formalized as the presence of *A* in the absence of *X*, whereas absence of awareness is formalized as the absence of *A* (with *X* either present or absent)[2]. The utility of this distinction between absence of awareness and awareness of absence is not limited to states of pure consciousness without intentional content (which may or may not exist, or be theoretically possible in principle; Metzinger, 2020), but also extends to modeling awareness of specific absences, such as being aware of the absence of a stimulus in a display.

Similar higher-order accounts of consciousness have posited that consciousness emerges from a higher-order representation of one's belief state and its relation with the external world (Lau, 2019), or a content-invariant representation in a higher-order state-space (Fleming, 2020). Like the Attention Schema Theory, these proposals dissociate between the specific contents of consciousness and being conscious of these contents (Lau & Rosenthal, 2011), and as a result allow for a representation of absence that is more than the absence of a representation. Some higher-order models have identified an abstract representation of presence vs. absence as the apex of a representational hierarchy (Fleming, 2020). Here we make a finer observation regarding the autonoetic nature of absence representations more specifically. The capacity of higher-order accounts of consciousness to model the distinction between awareness of absence and absence of awareness speaks to the idea that a representation of absence, more so than representations of presence, is in essence higher-order. The critical difference between the belief state *X is absent* and the absence of the belief state *X is present* is my counterfactual belief that I would have detected *X* had it been presented. This depends on my ability to represent myself and my mental states under varying conditions, a capacity that is not available to first-order models.

This distinction between higher-order and perceptual aspects of consciousness can be further formalized using more detailed computational models. For example, Lau (2007) proposed an extended SDT framework to explain how the phenomenon of blindsight can result from an improper placement of the decision criterion. Blindsight is the loss of subjective awareness following damage to the primary visual cortex despite relatively preserved discrimination performance. According to Lau's account, damage to the striate cortex results in a global decrease in the strength of evoked responses to visual stimuli. If subjects know about this

[2] Note that in AST, the critical content-invariant aspect of awareness *A* is not attention, but higher-order beliefs about one's own attentional state. Therefore, both *A* and *X* represent specific contents, but *X* corresponds to the content of consciousness, and *A* to meta-level content about one's own state that is itself not necessarily accessible to consciousness.

change, they can adjust their criterion to be more liberal and be more likely to report experiencing a stimulus even for relatively weak signals. But if subjects don't know about this change to the signal distribution, they will not update the position of their criterion and will report not being aware of stimuli, even when these stimuli are objectively registered by the visual system as measured with 2-alternative forced choice tasks (Ko & Lau, 2012).

Lau's original model was designed to explain a persistent change following brain damage, but a similar logic can be applied to transient fluctuations in the signal distributions and in subjects' meta-representation of this. For example, Mazor et al. (2020) modeled subjective confidence in detection decisions using an extended SDT model, where beliefs about fluctuations in overall attention states affect the expected signal strength, which in turn affects the placement of a decision criterion in a perceptual detection task (see Fig. 1, panel c). In more elaborate Bayesian models, meta-inference about current attentional states not only affects beliefs about world states, but also drives the active deployment of endogeneous attention (Sandved Smith et al., 2020). Common to these higher-order models of awareness is a hierarchical structure of beliefs, with higher-level beliefs about content-invariant aspects of awareness (for example, beliefs about attentional state or expected firing rate following a lesion) informing lower-level inference about external world states (for example, the presence or absence of a stimulus).

The expected precision or strength of sensory signals can be further conditioned on the interaction between attention and signal properties such as spatial location and modality. For example, when listening to a concert, we may represent our sensory precision as being high for auditory signals, but low for visual signals. Similarly, when focusing on the center of a computer screen, we may represent our sensory precision as high near the center of the screen, but low at the periphery. As a result, and perhaps counterintuitively, detection threshold may be lowered for visual stimuli in the first case, or for visual stimuli in the periphery in the second case. The rationale behind this adjustment of the detection threshold as a function of higher order beliefs about sensory precision is in essence the same, regardless of whether the threshold is adjusted globally or in a more constrained manner. In both cases, the positioning of the criterion contributes to the conscious percept in a way that is independent of the actual content of the percept itself. As such, we refer to this higher level of the hierarchy as content-invariant.

3 Using awareness of absence to disentangle content-specific and content-invariant aspects of consciousness

In the previous section we outlined two approaches to the modeling of awareness reports within the framework of Signal Detection Theory. Both first- and higher-order models treat awareness reports as the result of subpersonal inference about the most likely state of the world given noisy sensory evidence, but only in higher-order models is this inference informed by (and potentially informs) beliefs about content-invariant aspects of awareness. We showed that this hierarchical structure endows higher-order models with a capacity to express awareness of absence. We now turn to examine how judgments of awareness of absence can be used to differentiate between content-specific and content-invariant NCCs. To illustrate our proposal, we will start with an example, based on the experiment described in Pitts et al. (2012).

Imagine that you are a participant in a psychological experiment. You just finished the first phase of the experiment, where you were asked to identify an occasional dim disc target on a red ring while your brain activity was monitored using EEG. You found this part pretty demanding but your general feeling is that you did well on the task. You are now asked whether you noticed any patterns within an array of little white lines displayed within the ring (see Fig. 2). You haven't noticed any patterns, so you tick 'no', but this question makes you think that you might have missed something. In the second phase you are asked to perform the same task again, but this time you are more attentive to the little white lines and you notice a few squares and diamonds appear. When analyzing the data, the researchers will contrast neural responses to squares and diamonds in the first and the second phases and attribute any differences they find to your awareness of the white shapes.

In similar experimental designs an occipital-parietal EEG signal at around 200 milliseconds after shape appearance is typically more negative in those trials where participants notice the appearance of the shapes (known as Visual Awareness Negativity or VAN; Förster et al., 2020, see Fig. 2, lower panel). In our example, the researchers could confidently classify this neural signature, if found, as the neural signature of visual awareness. This is because, due to the careful task design, other explanations of a difference in activity can be eliminated: it can't be differences in task relevance (the shapes are irrelevant in both phases of the experiment), it can't be the need to report (in both phases no report is needed of a single shape), and it can't be the physical properties of the visual display (as they are matched). The only remaining explanation is that the activation is specific to awareness of the patterns emerging in the second phase while being absent in the first phase. But there are still two possible interpretations of this effect. One is that the contrast reveals brain activity that is correlated with the perceptual contents of conscious experience: geometric shapes compared to random line configurations, or more abstractly, presence of a stimulus object versus absence of a stimulus object. A different interpretation is that the contrast reveals brain activity that is correlated with content-invariant aspects of awareness that are not directly linked to any specific aspect of the display, but to the state of being aware of something. Such changes in content-invariant aspects of awareness may comprise shifts in attentional state and/or metacognitive beliefs about attentional state, regardless of specific perceptual contents.

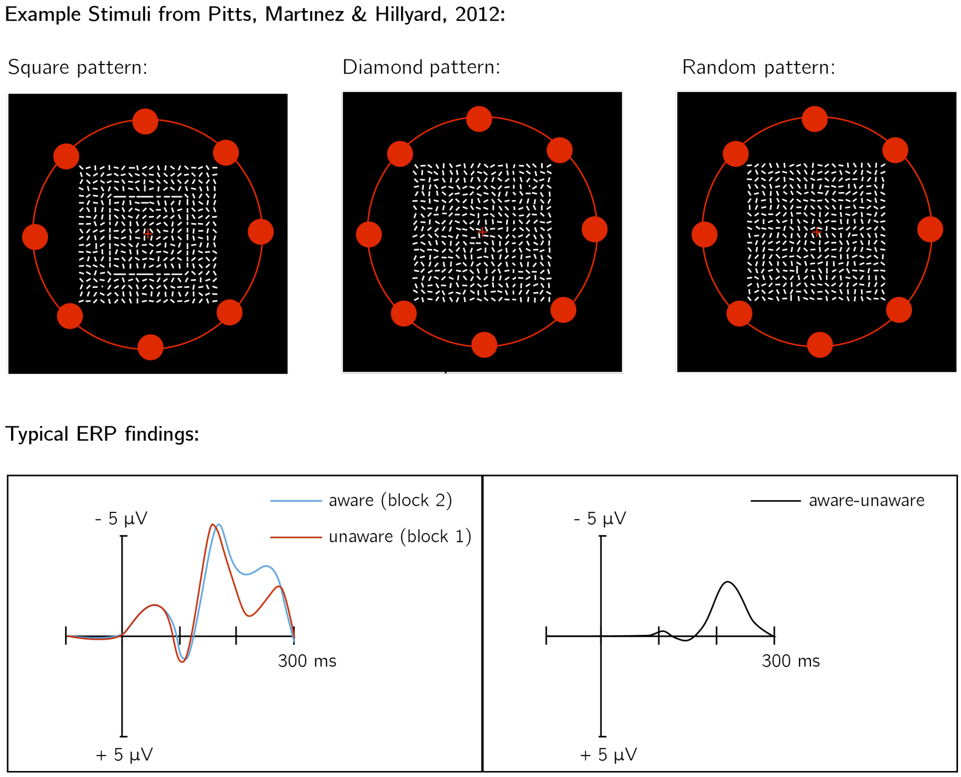

Figure 2: Upper panel:. Stimuli used in Pitts et al. (2012). In the original experiment, neural responses were recorded following the occasional appearance of a square or a diamond in the central display, as a function of awareness and task relevance. In our hypothetical inverted design, neural responses are recorded following the occasional disappearance of an existing square. Reproduced based on Figure 1, Pitts et al. (2012). Lower panel: Typical ERP over occipital electrodes in aware-unaware contrasts (illustration). Based on Figure 1, Förster, Koivisto, and Revonsuo (2020).

3.1 The inverted design

To make the distinction between these two alternatives clear, imagine a novel variant of Pitts' experiment. It begins just like the original one, except that the first display you see includes a diamond shape composed of a subset of the little white lines. After the first phase, you are asked whether you noticed the *disappearance* (rather than appearance) of the diamond at any point during the experiment. You haven't (you assumed the diamond was there all along), so you tick 'no', but this question makes you think that you might have missed something. In the second phase you are asked to perform the same task again, but this time you are more attentive to the little white lines and you notice that the diamond disappears occasionally for a brief time. What should we expect for the VAN component of the EEG in our alternative variant of the experiment? If the VAN correlates with content-invariant aspects of awareness – i.e. the state of being aware of the presence or absence of specific contents – this contrast should yield similar results to the ones in the original experiment, with a more negative ERP for Aware trials. However, if the VAN corresponds to content-specific aspects of awareness (for example, the appearance of a distinct stimulus in the visual array), we should expect a different pattern in the second experiment: the occipital-parietal signal at this time window should in fact be more similar to responses evoked by stimulus *offset* – stimulus disappearance – in the original experiment.

This hypothetical experiment is an example of what we refer to as an *Inverted Design*, in which participants actively report the disappearance of a previously visible target stimulus. Inverted designs resolve the content-specific/content-invariant ambiguity inherent in interpreting the aware-unaware contrast by incorporating awareness of stimulus absence as an experimental condition of interest. For the $\text{aware}_{\text{absence}}$–$\text{unaware}_{\text{absence}}$ contrast, a content-invariant NCC is predicted to show a similar activation profile to the more typical $\text{aware}_{\text{presence}}$–$\text{unaware}_{\text{presence}}$ contrast, whereas a content-specific NCC is predicted to show divergent results (because a content-specific NCC should track the content-specific awareness of stimulus presence). Importantly, this differential prediction for content-specific and content-invariant NCC is achieved without pharmacological or invasive interventions (in contrast with Aru et al., 2012; Klein & Barron, 2020), circumventing ethical concerns and practical challenges. Furthermore, using an inverted design allows detecting content-invariance in a different, potentially stronger sense than what is afforded by using a broad set of stimuli that spans multiple semantic categories (Rutiku et al., 2016). A neural marker of awareness that survives design inversion is invariant not only to various types of stimulus presence, but also to the more abstract notions of both stimulus presence and absence.

Inverting the typical awareness-of-presence design has the potential to advance our ability to tell apart these two classes of NCCs. However one concern is that reports of stimulus disappearance may conflate detecting stimulus absence and detecting the presence of the event of stimulus disappearance (for example, by relying on sensory transients). Such alternative accounts can be ruled out by introducing control conditions (for example, changes in the display that do not correspond with the appearance or disappearance of a stimulus), or by adopting gradual and illusory disappearances, where noticing the absence of a stimulus is not directly linked with any physical change to the display.

Recently, Davidson and colleagues (2020) combined a variant of this latter approach with a clever experimental manipulation to show that a particular neurophysiological signal is coupled to content-invariant rather than content-specific aspects of awareness. SSVEP (steady-state visually evoked potentials) reflect the entrainment of neural oscillations to rhythmic stimuli that can be recorded using EEG. In their design, participants were asked to report the illusory disappearance of peripheral targets, driven either by perceptual filling-in or by phenomenally matched physical disappearance. By allowing the background and targets to flicker at different frequencies, Davidson and colleagues could separately quantify SSVEPs induced by the target and background both before and during the disappearance event. Consistent with a content-invariant interpretation of the SSVEP, they found that SSVEPs for both target and background increased before the illusory disappearance of targets. If these SSVEPs were tracking the perceptual content of visual awareness, entrainment to the target frequency should have decreased before reports of disappearance, given that representations of specific content should presumably have become weaker when such content is absent. Instead, an increase in SSVEPs prior to target disappearance is in line with them tracking the focus of attention, but also more broadly with tracking content-invariant aspects of visual awareness. Increased alpha suppression (an independent marker of attentional effort) prior to events of illusory disappearance provided further support for this content-invariant interpretation. In summary, by inverting the A-U contrast and asking participants to report the disappearance of visual targets, Davidson and colleagues were able to decide between content-specific and content-invariant interpretations of a candidate NCC.

3.2 Two-dimensional report scheme

In typical experimental designs, awareness reports are given on a one-dimensional scale, from reports of minimal awareness of a stimulus, to full awareness of a stimulus. For instance, the commonly used “Perceptual Awareness Scale” (PAS) proposes four categories of experience going from “No experience”, to “Brief glimpse” to “Almost clear experience” to “Clear experience” (Ramsøy & Overgaard, 2004; Sandberg & Overgaard, 2015). As we have seen, the problem with such a scale is that it conflates both content-specific and content-invariant aspects of awareness: a report of “no experience” can be consistent with both absence of awareness and awareness of absence.

An alternative to such one-dimensional report schemes tackles this two-dimensional nature of awareness reports head-on, and allows participants to report both first-order perceptual contents, and their second-order assessment of awareness of these contents. This approach allows participants to report being fully aware of the absence of a stimulus, for example in the context of a visual detection task (see Fig. 3). We are not familiar with previous studies that have explicitly incorporated such a two-dimensional awareness report scheme. However, a related approach is to elicit visual detection judgments in conjunction with subjective confidence ratings (Kanai et al., 2010; Kellij et al., 2020; Mazor et al., 2020; Meuwese et al., 2014). Such a design allows subjects to report both high or low confidence in stimulus presence, and also high or low confidence in stimulus absence. These studies have consistently found generally lower confidence ratings for judgments about stimulus absence, even when these judgments are correct. Participants also exhibit poorer metacognitive ability to discriminate between missing an existing stimulus and correctly identifying the absence of that stimulus. This is consistent with an interpretation of ‘no’

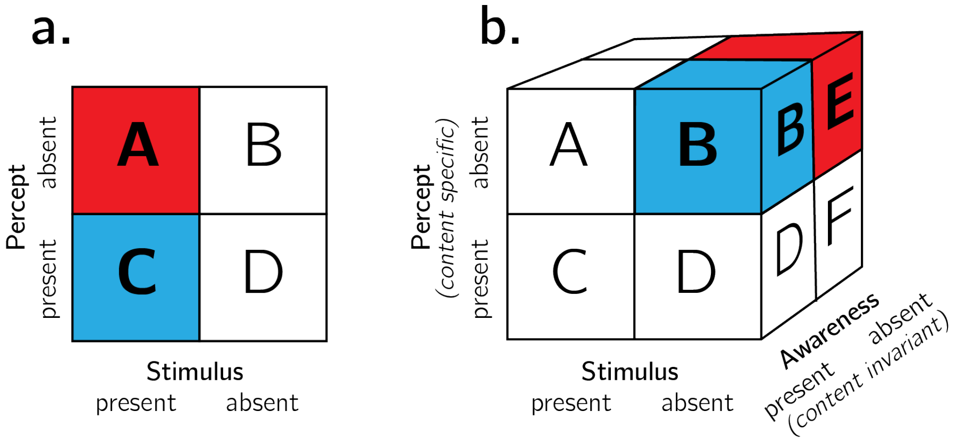

Figure 3: (a) In a typical aware-unaware contrast, trials in which participants perceived the stimulus are contrasted with trials in which they failed to perceive it. In this design both higher-order and perceptual states of awareness are interwined. (b) A two-dimensional report scheme allows a contrast of trials in which participants were aware of stimulus absence with trials in which they were unaware of stimulus presence (or absence).

responses as mostly reflecting absence of awareness of a stimulus, rather than (accurate) awareness of absence.

There was, however, one intriguing exception to this pattern. When stimulus visibility was manipulated using attentional manipulations such as the attentional blink or spatial uncertainty, rather than manipulations of the stimulus itself, Kanai et al. (2010) found that participants were more confident in correct-rejection trials compared to misses (and as a result had better metacognitive sensitivity for reporting absence). One interpretation of this effect is that with their attention fluctuating, participants were better able to know when they were likely to have missed the presence of a target, and therefore also whether they were likely to have correctly perceived stimulus absence. In other words, these attentional manipulations enhance the dependence of stimulus visibility on content-invariant aspects of consciousness, such that they begin to influence behavioural reports. This focus on awareness of absence under disturbed attention neatly reveals a potential contribution of beliefs about one's attention to the subjective content of awareness. The use of similar paradigms in conjunction with neuroimaging may open up new avenues for dissociating content-specific and content-invariant NCCs.

4 A neurofunctional separation between content-invariance and content-specificity

Which NCCs are sensitive to the presence or absence of an actor, and which to the presence or absence of a spotlight? Based on currently available data, we can

Mazor, M., & Fleming, S. M. (2020). Distinguishing absence of awareness from awareness of absence. *Philosophy and the Mind Sciences*, *1*(II), 6. https://doi.org/10.33735/phimisci.2020.II.69

make some crude guesses. Sensory regions are likely to represent local contents of awareness, such as subpersonal beliefs about stimulus presence or absence (Boly et al., 2017; Koch et al., 2016). In contrast, lateral and anterior aspects of prefrontal cortex may track higher-order, global aspects of conscious states, and have been associated both with reality monitoring and criterion-setting (Del Cul et al., 2009; Lau & Rosenthal, 2011; Simons et al., 2017; Vallesi, 2012). More specifically, associations between activation in the lateral prefrontal cortex and subjective awareness may reflect beliefs about global states of attention and the translation of these beliefs into policy changes (such as setting the criterion for awareness reports), rather than beliefs about the presence of a specific target stimulus.

Consistent with a role for these regions in attention monitoring and criterion setting, activation in parietal and prefrontal cortical areas have been shown to decrease not only in trials where participants reported a stimulus to be highly visible, but also in trials where participants reported the stimulus not to be visible at all, compared to intermediate subjective visibility ratings (Binder et al., 2017; Christensen et al., 2006). This nonmonotonic effect may plausibly reflect a projection of a monotonic effect of a higher-order construct (such as beliefs about sensory precision or subjective confidence) onto a content-specific 'visibility' dimension. For example, trials in which participants report extremely low or high visibility ratings may be similar in that in both cases participants believe they were attentive and expected high fidelity sensory signals (high precision, in the parlance of predictive coding models) as a result. Indeed, when given the option to report stimulus presence and absence independently from subjective confidence, modulation of prefrontal activation by confidence was highly similar for 'yes' and 'no' responses in a perceptual detection task (Mazor et al., 2020). A role for the prefrontal cortex in higher-order monitoring of internal states is also consistent with relative increases in prefrontal activation in paradigms that require explicit report, compared to no-report paradigms (Frassle et al., 2014). The hypothesis that the lateral prefrontal cortex is representing these higher-order aspects of awareness rather than content-specific percepts can be directly tested using an inverted design or a two-dimensional report scheme similar to the ones we discuss above.

5 Conclusion

A popular approach in the quest for the NCC has been to contrast aware and unaware trials during near-threshold perception. However, such an Aware-Unaware contrast conflates neural correlates of transient fluctuations in content-invariant aspects of consciousness with neural correlates of the specific perceptual contents of consciousness. This problem is most pertinent in trials where participants report being unaware of a stimulus, as such trials can reflect either awareness of absence or absence of awareness. Here we advocate for the adoption of theoretical frameworks that make this distinction explicit, such as higher-order or multi-level models. To map the different levels of these models to neurophysiological markers, we suggest the use of tasks in which beliefs about stimulus presence and absence are manipulated independently of reports of awareness.

Acknowledgments
We thank the two anonymous reviewers for their illuminating comments and suggestions. This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust. S.M.F. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206648/Z/17/Z). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

References

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Aru, J., Suzuki, M., Rutiku, R., Larkum, M. E., & Bachmann, T. (2019). Coupling the state and contents of consciousness. *Frontiers in Systems Neuroscience*, *13*, 43. https://doi.org/10.3389/fnsys.2019.00043

Baars, B. J. (1993). *A cognitive theory of consciousness.* Cambridge University Press.

Bayne, T., & Hohwy, J. (2013). Consciousness: Theoretical approaches. In A. E. Cavanna, A. Nani, H. Blumenfeld, & S. Laureys (Eds.), *Neuroimaging of Consciousness* (pp. 23–35). Springer Berlin Heidelberg.

Binder, M., Gociewicz, K., Windey, B., Koculak, M., Finc, K., Nikadon, J., Derda, M., & Cleeremans, A. (2017). The levels of perceptual processing and the neural correlates of increasing subjective visibility. *Consciousness and Cognition*, *55*, 106–125. https://doi.org/10.1016/j.concog.2017.07.010

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and Neuroimaging Evidence. *The Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Christensen, M. S., Ramsøy, T. Z., Lund, T. E., Madsen, K. H., & Rowe, J. B. (2006). An fMRI study of the neural correlates of graded visual perception. *NeuroImage*, *31*(4), 1711–1725. https://doi.org/10.1016/j.neuroimage.2006.02.023

Davidson, M. J., Mithen, W., Hogendoorn, H., van Boxtel, J. J., & Tsuchiya, N. (2020). The SSVEP tracks attention, not consciousness, during perceptual filling-in. *eLife*, *9*, e60031. https://doi.org/10.7554/eLife.60031

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, *100*(14), 8520–8525. https://doi.org/10.1073/pnas.1332574100

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, *5*(10), e260. https://doi.org/10.1371/journal.pbio.0050260

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*(9), 2531–2540. https://doi.org/10.1093/brain/awp111

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *2020*(1), niz020. https://doi.org/10.1093/nc/niz020

Förster, J., Koivisto, M., & Revonsuo, A. (2020). ERP and MEG correlates of visual consciousness: The second decade. *Consciousness and Cognition*, *80*, 102917. https://doi.org/10.1016/j.concog.2020.102917

Frassle, S., Sommer, J., Jansen, A., Naber, M., & Einhauser, W. (2014). Binocular rivalry: Frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience*, *34*(5), 1738–1747. https://doi.org/10.1523/JNEUROSCI.4403-13.2014

Graziano, M. S. A. (2013). *Consciousness and the Social Brain.* Oxford University Press.

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *06*. https://doi.org/10.3389/fpsyg.2015.00500

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* John Wiley.

Helmholtz, H. von. (1948). Concerning the perceptions in general, 1867. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 214–230). Appleton-Century-Crofts.

Kanai, R., Walsh, V., & Tseng, C.-h. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057. https://doi.org/10.1016/j.concog.2010.06.003

Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A. K., & Odegaard, B. (2020). An investigation of how relative precision of target encoding influences metacognitive performance. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-020-02190-0

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130204. https://doi.org/10.1098/rstb.2013.0204

Klein, C., & Barron, A. B. (2020). How experimental neuroscientists can fix the hard problem of consciousness. *Neuroscience of Consciousness*, *2020*(1), niaa009. https://doi.org/10.1093/nc/niaa009

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1401–1411. https://doi.org/10.1098/rstb.2011.0380

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv*. https://doi.org/10.31234/osf.io/ckbyf

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. In *Progress in Brain Research* (Vol. 168, pp. 35–48). Elsevier. https://doi.org/10.1016/S0079-6123(07)68004-2

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763–18768. https://doi.org/10.1073/pnas.0607716103

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Mazor, M., Friston, K. J., & Fleming, S. M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *eLife*, *9*, e53900. https://doi.org/10.7554/eLife.53900

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of “pure” consciousness. *Philosophy and the Mind Sciences*, *1*(I), 1–44. https://doi.org/10.33735/phimisci.2020.I.46

Meuwese, J. D. I., Loon, A. M. van, Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-014-0643-1

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *The Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1523/JNEUROSCI.3217-16.2017

Pitts, M. A., Martínez, A., & Hillyard, S. A. (2012). Visual processing of contour patterns under conditions of inattentional blindness. *Journal of Cognitive Neuroscience*, *24*(2), 287–303. https://doi.org/10.1162/jocn_a_00111

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Rutiku, R., Aru, J., & Bachmann, T. (2016). General markers of conscious visual perception and their timing. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00023

Sandberg, K., & Overgaard, M. (2015). Using the perceptual awareness scale (PAS). In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 181–196). Oxford University Press.

Sandved Smith, L., Hesp, C., Lutz, A., Mattout, J., Friston, K., & Ramstead, M. (2020). Towards a formal neurophenomenology of metacognition: Modelling meta-awareness, mental action, and attentional control with deep active inference. *PsyArXiv*. https://doi.org/10.31234/osf.io/5jh3c

Simons, J. S., Garrison, J. R., & Johnson, M. K. (2017). Brain Mechanisms of Reality Monitoring. *Trends in Cognitive Sciences*, *21*(6), 462–473. https://doi.org/10.1016/j.tics.2017.03.012

Vallesi, A. (2012). Organisation of executive functions: Hemispheric asymmetries. *Journal of Cognitive Psychology*, *24*(4), 367–386. https://doi.org/10.1080/20445911.2012.678992