Introspection Is Signal Detection

Jorge Morales

Introspection is a fundamental part of our mental lives. Nevertheless, its reliability and its underlying cognitive architecture have been widely disputed. Here, I propose a principled way to model introspection. By using time-tested principles from signal detection theory (SDT) and extrapolating them from perception to introspection, I offer a new framework for an introspective signal detection theory (iSDT). In SDT, the reliability of perceptual judgements is a function of the strength of an internal perceptual response (signal-to-noise ratio), which is, to a large extent, driven by the intensity of the stimulus. In parallel to perception, iSDT models the reliability of introspective judgements as a function of the strength of an internal introspective response (signal-to-noise ratio), which is, to a large extent, driven by the intensity of conscious experiences. Thus, by modelling introspection after perception, iSDT can calibrate introspection's reliability across a whole range of contexts. iSDT offers a novel, illuminating way of thinking about introspection and the cognitive processes that support it.

1. Introduction

The study of introspection has a thorny history. Introspection has been praised as an infallible capacity, vilified as utterly unreliable, and everything else in between. How can this be? How can there be such a dispute about the trustworthiness of one of our most important capacities? To make progress around these disputes, a successful theory of introspection should aim to calibrate its whole range of operation and explain its reliability conditions: when and why it succeeds and when and why it fails.

My goal here is to provide a new framework for explicating and calibrating introspection. To do so, I will take conceptual and theoretical insights from the science of perception—in particular, from signal detection theory (SDT)—and extrapolate them to model first-personal access to conscious experiences as a type of signal detection. I call the result of this model extension 'introspective signal detection theory' (iSDT). The main aim of introducing the iSDT framework is to help us conceptualize introspection in a more systematic way than previous approaches have typically allowed. This new theoretical apparatus can handle a wider range of cases (both successes and failures) by appealing to a single machinery whose fundamental

Electronically published April 5, 2024.

The British Journal for the Philosophy of Science, volume 75, number 1, March 2024.
 https://doi.org/10.1086/715184

underlying operation is shared by other cognitive capacities (perception, memory, and decision-making, among others). In addition to calibrating introspection's whole range of reliability, iSDT also offers a unifying way of understanding response bias and confidence in introspective judgements. In brief, by offering a functional analysis of the psychological principles under which introspection operates, a fruitful sketch of how to model its reliability and the computational mechanisms that support it will emerge.

Science can make progress by applying familiar, well-understood concepts and models from one domain to unfamiliar, less well-understood concepts and models in a different domain. This phenomenon is known as 'model template transfer' (Humphreys [2002]; Knuuttila and Loettgers [2016]) or, more generally, 'model migration' (Lin [unpublished]). Beyond mathematical and computational structures, model templates are useful for the conceptual resources they provide; in fact, model templates 'enable cross-disciplinary transfer, sensitizing us to perceive similar patterns across wide variety of different kinds of empirical systems [. . .] they offer resources for further investigation and new theoretical insights' (Knuuttila and Loettgers [2014], p. 298).

One of the most successful cases of model migration is Maxwell's ([1861]) successful transformation of Faraday's mechanical model of fluids to explain electromagnetic fields. Recent examples include the extension of game-theoretic models to evolutionary decision-making (Smith and Price [1973]) or the extension of tools developed for understanding the random motion of suspended particles for modelling financial markets (Merton [1969]; Samuelson [1969]). In psychology, the most influential model extension is, without a doubt, signal detection theory. Originally developed during the first half of the twentieth century as a mathematical framework for evaluating radar performance, SDT was later adapted to explain perception (Tanner [1954]; Macmillan and Creelman [2005]). Since then, SDT has been described as 'one of psychology's most well-known and influential theoretical frameworks' (Wixted [2020], p. 201) and even as 'the most towering achievement of basic psychological research of the last half century' (Estes [2002], p. 15). By taking SDT's insights and conceptual apparatus to model introspection, we can make progress in a domain that has historically resisted satisfactory modelling in both philosophy and psychology.[1]

In a nutshell, SDT models perception as the joint outcome of perceptual discrimination and decision-making. An observer's ability to discriminate stimuli (called perceptual sensitivity) depends on the strength of the perceptual evidence their visual system accumulates. In turn, the perceptual evidence (also known as the internal perceptual response or perceptual response, for short) tends to be proportional to the strength of the stimulus itself. In addition to discriminating the stimulus, when

[1] Signal detection in general, and SDT in particular, has been successfully applied outside the perceptual domain. For a recent elegant use of SDT to model memory, see (Schurgin et al. [2020]).

making a perceptual judgement, observers make a decision to classify the perceptual evidence in one way or another. To do so, they set a criterion or response rule (also known as response bias) to determine the level of internal response required for classifying the stimulus. Consider the following simple scenario: everything else being equal, an observer is more likely to perceive accurately a person in an alley when the alley is well-lit (strong stimulus, strong internal response) than when the person is in a dark alley (weak stimulus, weak internal response). Moreover, whether the same amount of perceptual evidence leads to discriminating or not someone in the alley depends on how liberal or conservative the observer's criterion is.

The view I introduce here—iSDT—models introspection similarly to how SDT models perception. Accordingly, iSDT models introspective judgements as the joint outcome of an introspective discrimination and a decision. The central tenet of iSDT is that the intensity of our conscious experiences (what I call 'mental strength') modulates the internal introspective response (or introspective response, for short), and this, in turn, modulates introspective sensitivity. Thus, iSDT proposes that everything else being equal, an introspector is more likely to introspect accurately an intense experience (a strong pain, a vivid mental image, and so on) than a weak experience (a weak pain, a faint mental image, and so on). iSDT also relies on introspective response bias to fully account for introspective judgements. For example, for an identical weak experience, a liberal introspector may judge they are undergoing that experience (a weak pain, a faint mental image, and so on) while a conservative introspector may not.

In section 2, I discuss desiderata for calibrating introspection as well as iSDT's most basic assumptions about the nature of introspection. I also discuss iSDT in the context of other theories of introspection, in particular other inner-sense theories. In section 3, I offer an overview of SDT, especially notions such as sensitivity, response bias, and confidence. Section 4 introduces the notion of 'mental strength' (conscious experience intensity) and discusses its connection, on one hand, to stimulus intensity and perceptual response, and, on the other hand, to introspective response. In section 5, I develop the iSDT framework and focus on the introspection of pains as a case study. Finally, in section 6, I discuss generalizations of iSDT to introspection of mental imagery and perceptual experiences. Furthermore, I show how iSDT offers systematic explanations for different types of empirical results taken from the scientific study of consciousness.

2. Introspection as a Garden-Variety Capacity

The problem of calibration 'arises for any scientific instrument and cognitive capacity' (Goldman [2004], p. 14). Introspection is no exception. Calibrating introspection requires determining when it works and when it does not. Two desiderata for an adequate theory of introspection emerge. First, it should have the right scope, that is, it should explain introspection's full range of reliability. This means that the conditions

that favour both accurate and inaccurate introspective judgements should be covered by the theory. Second, a theory of introspection must be illuminating. This means that the theory not only should cover the whole range of relevant cases, but also should explain why introspection has the range of reliability that it has. Everything else being equal, it is desirable that this explanation is the same (or of the same kind) for successes and failures.[2]

These desiderata apply to the calibration of other faculties too. SDT is a successful example of a theory that explains perception's full range of reliability in an illuminating way. SDT explains perceptual sensitivity by appealing to the signal-to-noise ratio of the internal perceptual response, thus covering perceptual sensitivity's whole range—from chance to performance at ceiling. By appealing to this single principle, SDT can explain (and predict!) why perception is good when it is good and why it is bad when it is bad. Similarly, a theory of introspection should explain (and predict) when accurate and inaccurate introspection is likely to happen by appealing to a unified principle. iSDT is such a theory.

Prima facie, a reasonable assumption when thinking about introspection's reliability is that introspection is not unlike the rest of our cognitive capacities.[3] Call this the assumption that introspection is a garden-variety capacity. Part of what it means to be a garden-variety capacity is that it is not equally reliable in all conditions. Like any of our other faculties, introspection may sometimes get things right, and it may sometimes get things wrong.

As I understand it here, introspection is the focusing of one's attention on one's current conscious experiences to make judgements about them.[4] Accordingly, we can introspect pains, mental images, perceptual experiences, and emotions, among others. Introspection thus understood implies some amount of effort from the introspector (for example, directing their attention in the right time and manner). Thus, introspective judgements are a kind of cognitive achievement susceptible to success and failure (and this is true even if the effort is minimal). An implication of this way of understanding introspection is that, at least sometimes, we undergo conscious experiences that we do not (fully) introspect. For example, one does not always direct attention towards, and make introspective judgements about, peripheral vision.[5] In any case, iSDT will try to capture these types of introspective judgements. Relatedly,

[2] Failure to meet the *ceteris paribus* clause would make room for pluralist accounts of introspection (Schwitzgebel [2012]; Renero [2019]).

[3] Naturally, some philosophers conclude that introspection is special. Here I just suggest that assuming introspection is not special is a reasonable starting point.

[4] Many philosophers agree that introspection involves some kind of attention oriented towards conscious experiences. Note that they agree despite espousing very different views about introspection (and the mind). To cite just a few: (Peacocke [1998]; Carruthers [2000]; Hatfield [2005]; Rosenthal [2005]; Goldman [2006]; Ryle [2009]; Chalmers [2010]; Schwitzgebel [2012]; Wu [2014]; Giustina and Kriegel [2017]).

[5] Strictly speaking, a view where introspection and consciousness are not independent could still embrace this way of understanding introspection. For example, someone who holds that all conscious states are introspected could still agree that one introspect more (and hence is more conscious of) the centre of the visual field than the periphery.

a theory aiming to calibrate introspection need not depend on a specific theory of consciousness, and this is true of the theory I develop here. Finally, for reasons of space, I will focus only on the introspection of conscious sensory experiences such as pain, mental images, and perceptual experiences.[6]

2.1. Infallibility and unreliability

The garden-variety assumption and this way of understanding introspective judgements are in clear tension with prominent views that take introspection to be infallible (self-intimating, transparent, privileged, or impervious to error in some other way). A similar tension is present too for views that consider introspection to be completely unreliable. Both kinds of approach bypass the problem of calibration: if introspection is always or never to be trusted, there is no range of operation to be established. Here, I briefly comment on these positions.

Views that consider introspection to be infallible have a long history. Descartes ([1984], AT VII 29), for example, vividly evokes introspective infallibility when he writes: 'I am now seeing light, hearing a noise, feeling heat. But I am asleep, so all this is false. Yet I certainly seem to see, to hear, and to be warmed. This cannot be false'. More recently, Gertler ([2001], p. 321) argues that introspection takes place via pure demonstrative reference achieved by directing attention to the phenomenal contents of our conscious experiences: 'By appropriately attending to the dull throbbing sensation [of a headache], you demonstratively pick out the phenomenal content <dull throbbing>'. Phenomenal contents are supposed to be embedded in the introspective judgement 'it is thus here and now', thus preventing any sort of error when introspecting one's conscious experiences. Gertler's understanding of introspection explicitly denies the garden-variety assumption, since she thinks introspection works differently from any other mental mechanism. Introspectors 'grasp the content directly [. . .] in the sense that there is no causal gap between the referring state and its referent, the phenomenal content' (Gertler [2001], p. 232). Several other defences of some sort of introspective infallibility—especially about occurrent phenomenal experiences—abound in the recent literature (Shoemaker [1996]; Chalmers [2003]; Horgan and Kriegel [2007]).

My goal here is not to discuss these views at length. But I do want to highlight that introspective infallibility is often defended based on a very limited set of examples. It might be tempting to think introspective judgements are always accurate if the examples one relies on are of the type 'I'm in pain now', 'I am seeing a red patch', or 'I'm experiencing this'. As Schwitzgebel ([2008], pp. 259–60) correctly points out, 'there is a reason optimists like the example of pain and foveal visual experience of a single bright colour. It is hard, seemingly, to go too badly wrong in introspecting

[6] In principle, iSDT could be extended to other conscious experiences as long as they have an intensity dimension.

really vivid, canonical pains and foveal colours. But to use these cases only as one's inference base rigs the game'. Once more complex (yet completely common) cases are considered, introspection's infallibility becomes much harder to maintain.

This acute observation about this 'diet' of examples, however, need not turn us into sceptics about introspection. For instance, Schwitzgebel ([2008], p. 247) thinks that 'we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or "phenomenology")'. Rather than embracing this equally extreme position, what we need is a principled method for calibrating the whole range of reliability of our capacity to introspect. By taking the garden-variety assumption as our starting point, we should find it equally implausible that introspection is infallible and that it is always utterly broken. Just as we try to understand why perception, memory, decision-making, and other cognitive capacities work when they do and why they fail when they do, we should find systematic ways to model introspection's range of operation. In any case, this will be my goal here.

2.2. Inner sense

By departing from the tradition that considers introspection infallible, iSDT can also abandon a tradition that considers introspection unique or special. Rather, iSDT takes introspection to function similarly to other faculties—perception in particular—and thus embraces a tradition that treats introspection as a kind of 'inner sense'.[7] Theories of inner sense have many prominent defenders (Armstrong [1968]; Locke [1975]; Lycan [1996]; Kant [1998]; Goldman [2006]). Nevertheless, this kind of theory has acquired a bad reputation. So bad that philosophers often find appeals to inner sense 'unpersuasive, even repugnant' (Goldman [2006], p. 225). Against this trend, iSDT aims to become an attractive option for modelling introspection.

Part of the distaste for inner-sense mechanisms stems from common simplifications by critics and, sometimes, champions too. Armstrong ([1968], p. 325), a notable proponent of an inner sense, compares introspection to bodily perception because it happens without a 'proper organ' and its object 'is private to each perceiver'. While it is true that introspection does not have a proper organ, the comparison is unfortunate. Critics of inner sense sometimes also base their objections on misguided analogies. Hill ([1988], pp. 12–13), for example, writes that an inner 'scanning device is said to stand in much the same relationship to sensations as the physical eye does to extramental objects and events'. Neither of these, however, is an adequate point of similarity between perception and introspection. Rather than a literal organ (or lack thereof), it is the type of internal processing they carry out that makes perception and introspection similar.

[7] Others also abandon the infallibility claim or the uniqueness claim or both. To list just a handful of recent examples, see (Rosenthal [2005]; Schwitzgebel [2008]; Bayne and Spener [2010]; Hohwy [2011]; Reuter [2011]; Giustina and Kriegel [2017]).

Hill raises an interesting criticism against inner-sense views. According to him, the inner-sense analogy gives the wrong result: while the internal qualities of extramental entities 'are never affected by their coming to stand in [any informational relation to the physical eye]', defenders of an inner-scanning mechanism cannot argue that 'the internal qualities of sensations do not change when one scans them' (Hill [1988], p. 13). But inner-sense theorists need not deny that the scanning mechanism alters its target states, nor do they need to accept that perceptual processing does not alter its target.[8]

On one hand, it is not generally true that detection mechanisms do not alter their target objects. Measuring an object's temperature without altering it—even if just marginally—is practically impossible. So, when detecting our experiences, we alter them. For instance, introspecting may make experiences stronger: a pain may become more intense, a mental image may become more vivid, a visual experience may become more striking (see sec. 4 for further discussion of this point). This implies that we hardly, if ever, introspect 'pure' experiences. This is, of course, the right result (one that Hill himself embraces): we cannot know exactly what an unintrospected (that is, an unattended) conscious experience is like. How could we if we are not introspecting it! If we wanted to say something about 'pure' experiences, we would need to rely on memory of an unintrospected experience. It should be obvious that this opens the possibility of significant error.

On the other hand, the claim that the eye does not alter the internal qualities of its objects is somewhat misleading. While perhaps literally true, the right comparison between perception and introspection is not between the eye and some internal organ. Rather, it is between perceptual and introspective processing. Introspection can be successfully modelled after perception, but introspection is not perception any more than perception is receiving radio signals, the original domain of application of SDT models. Moreover, orienting our eyes (foveating) and, more importantly, orienting our attention towards an object most definitely alter the perceptual representation and perceptual experience of seeing that object (Carrasco et al. [2004]; Carrasco [2011]). In summary, the inner-sense theorist can admit that introspective attention affects target experiences, but the critic must admit that a fair comparison with perception would highlight that perceptual attention also alters perceptual processing of the target stimulus.

Shoemaker's ([1996], p. 275) criticism of inner sense is also worth considering here. Self-blindness occurs when a creature capable of conceiving certain kind of mental facts and phenomena is, nevertheless, incapable of gaining introspective access to such mental facts and phenomena: 'He is in extreme pain, his pains are extremely unpleasant, but there is nothing bad about this because he is unaware of his pains [. . .] His pains hurt, but they do not hurt him'. Self-blindness, according to Shoemaker, is 'nonsense'. Instead, he defends self-transparency, which holds that,

[8] Incidentally, there are several points of agreement between iSDT and Hill's positive proposal. For example, he does not think we introspect every conscious experience and he allows introspective errors.

necessarily, if you are in a mental state M, and various background conditions obtain, and you are rational, you will believe you are in M.[9]

Inner-sense approaches to introspection deny self-transparency and embrace the possibility of self-blindness. In particular, inner-sense views assume that target conscious states exist independently from the subject becoming (accurately) aware of them introspectively (for example, if the detection mechanism were absent, inoperant, or otherwise faulty).

As before, the goal here is not to offer an in-depth analysis and rebuttal of Shoemaker's view (among many others, those who have offered convincing arguments against self-transparency include Hill [1988]; Williamson [2002]; Srinivasan [2015]). Rather, I aim to contextualize some of iSDT's assumptions. Contrary to Shoemaker's suggestions, not only some kind of self-blindness is empirically possible, but we should also take the intensity of the introspected experiences into account in our explanations of introspection.

Recent evidence shows that perceptual states can remain intact while self-reflective mechanisms are corrupted. In cases of metacognitive failure, subjects display normal performance in a perceptual task at the same time that they display altered self-evaluations of their performance in those tasks. These effects have been observed in both neuropsychological populations (Fleming et al. [2014]), via causal interventions in neurotypical subjects (Rounis et al. [2010]; Cortese et al. [2016]) and via psychophysical manipulations (Zylberberg et al. [2012]; Koizumi et al. [2015]; Samaha et al. [2016]; Morales et al. [2022]; Maniscalco et al. [unpublished]). While these effects do not show complete self-blindness, some of them are not subtle either. For example, Fleming and colleagues ([2014]) showed a 50% reduction in metacognitive efficiency in patients with prefrontal cortex lesions. Importantly, these patients saw the stimuli without any trouble and in a completely normal way as revealed by their ability to do the task. As this case suggests, metacognitive self-reflecting mechanisms can fail while perceptual experiences remain intact.

The metacognitive ability to assess one's performance in a perceptual task is technically not identical to introspecting the experiences one undergoes while doing said task. However, the differences are not relevant in practice (Morales and Lau [2022]). Asking subjects to rate confidence in their performance produces virtually identical results to asking subjects to introspect how visible the stimulus was (Peters and Lau [2015]). Even though a confidence report may not be a perfect substitute for an introspective report about ongoing phenomenology, the subjective feeling of having perceived a stimulus is partially supported by our introspective ability to know our own experiences. And this is especially true in introspective-reliant tasks, that is, tasks that demand focusing on the quality of one's experience (as opposed to focusing on the stimuli themselves) (Chirimuuta [2014]; Spener [2015]). Thus, although distinct in principle, metacognitive failure in fact provides a window into introspective failure.

[9] Here I follow Stoljar's ([2022]) reconstruction of Shoemaker's self-transparency claim.

Beyond potential malfunctions of the introspective apparatus, reliability under completely normal circumstances is not constant across conditions. Weak pains (or faint mental images or weak perceptual experiences) and strong pains (or vivid mental images or intense perceptual experiences) are not introspectable with equal accuracy. Blank statements such as self-transparency lack crucial information about the intensity of the mental state and, therefore, cannot be appropriately evaluated in the iSDT framework (or any framework that accepts the garden-variety assumption). Even relaxing the modal claim in self-transparency by substituting 'necessarily' for something weaker such as 'in normal cases' or 'most of the time' or even 'ideally' is not sufficient. The lack of details about the intensity of the experience remains problematic. For example, if mental state M is substituted for 'a very weak pain', iSDT would not predict that 'most of the time' or even 'ideally' you would believe that you are in (a weak) pain. In contrast, iSDT would predict that 'most of the time' or 'ideally' you acquire such a belief when in 'extreme pain' (which is closer to Shoemaker's example cited above). But more importantly, this shows that iSDT makes distinct predictions about the reliability of introspection depending on the degree of intensity of the targeted experience.

In the next sections, I will develop the building blocks of the iSDT framework, which will allow us to think about introspection in a systematic way. The framework has a wide scope (it explains success and failure), and it is explanatorily illuminating (it explains why these cases succeed and fail, and it does so by appealing to a single kind of mechanism). Moreover, the framework achieves this with the minimal assumption that introspection is a garden-variety capacity and that, thereby, it operates in a similar fashion to the rest of our cognitive capacities—perception in particular.

3. Signal Detection Theory

Consider the next scenarios. They all assume that a man is in an alley and that his face is in your direct line of sight.

Bright Alley: You walk by an alley late at night. The alley's lamp is on, so it is easy for you to see a man next to a refuse can. His face looks bright, and the contours of his facial features look well defined. You are confident you are seeing someone.
Dark Alley: The alley's lamp is off. The man's face looks dark, and the contours of his face look ill defined. It is hard for you to see him. You mistakenly categorize his face as being just a shadow and judge that the alley is empty. However, you are not confident.
Dark Alley + News: Identical situation to the dark alley scenario except that you heard that a robber is on the run in the neighbourhood. You categorize the ill-defined shadow as someone's face. Note that the man's face in the dark is visually processed in exactly the same way it is processed in the dark alley. The only difference here is that knowing about the robber changes how you categorize the same evidence,

thereby changing your perceptual judgement. You are still not confident about what you see.

These scenarios illustrate three paradigmatic features of perception that SDT can model successfully: sensitivity, response bias, and confidence. In a nutshell, according to SDT, perceptual judgements are determined by sensory sensitivity (namely, the ability to discriminate stimuli based on the way these shape a psychological decision space) and by response biases (namely, the manner in which the psychological space is partitioned to generate possible responses) (Macmillan and Creelman [2005]).

3.1. Sensitivity

Paradigmatically, perception is modelled as an observer deciding whether an internal perceptual response *p* was generated by a stimulus class S1 (for example, stimulus absent, line oriented left, and so on) or S2 (for example, stimulus present, line oriented right, and so on). The perceptual response corresponds to the strength of sensory evidence, in turn modulated by the intensity of the stimulus. A fundamental assumption of SDT is that across repeated presentations of the same stimulus class, the perceptual response can have different values due to ever-present random noise (either in the stimulus or in perceptual processing). The dimension along which the values of the internal response are distributed is called 'the decision axis'. The perceptual response *p* in any given case can be thought of as being drawn from either a noise or a signal-plus-noise distribution (fig. 1).

3.2. Response bias

Because the distributions overlap, it is always possible for a given value of *p* to have been generated by S1 or S2. To make a perceptual judgement, observers classify *p* as S2 if it exceeds a response criterion *c* (solid lines in figs. 1 and 2), and as S1 otherwise. Importantly, whereas sensitivity is a function of stimulus properties and perceptual processing (typically) beyond the observer's control, *c* reflects a response strategy largely determined by the observer's priors, preferences, goals, and other traits (for example, maximizing the probability of responding correctly, maximizing rewards, degree of risk aversion, perceptual biases, and so on).

Importantly, as the dark alley and the dark alley + news scenarios illustrate, sensitivity and response bias are independent from each other (figs. 2B and 2C). While preserving identical sensitivity (the distance between the distributions' means is the same), an identical perceptual response can yield different perceptual judgements due to changes in the detection criterion. In the dark alley + news scenario, the criterion for detecting the presence of someone lurking in the alley becomes more liberal (fig. 2C). This criterion change results in changes in response accuracy in these trials (even if overall sensitivity remains the same): a correct classification in the dark alley + news scenario (hit) and an incorrect classification in the dark alley scenario (miss).

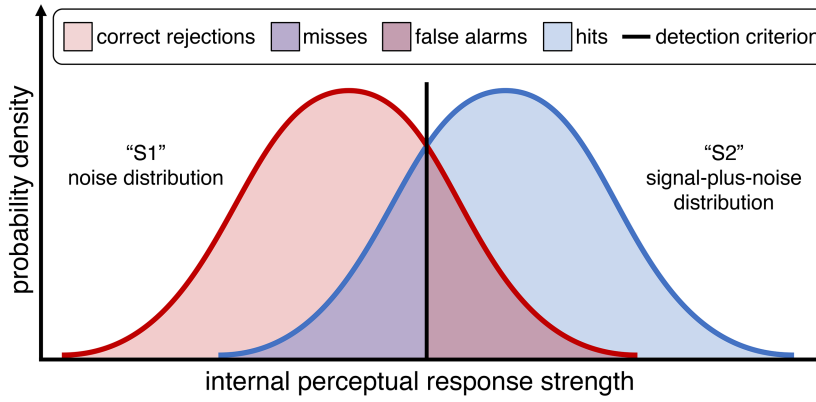

Figure 1. Signal detection theory model. SDT models perception as a decision about perceptual internal evidence. Observers decide whether an evidence sample was drawn from class S1 (noise distribution) or S2 (signal-plus-noise distribution). Perceptual judgements depend on whether the drawn sample is below or above a decision criterion (solid vertical line). If the sample is below the criterion, the observer selects S1; if it is above the criterion, the observer selects S2. There are two kinds of correct trials. Correct rejections take place when the observer selects S1 when the sample is drawn from the S1 distribution; selecting S2 when a sample is drawn from the S2 distribution is called a hit. There are two kinds of incorrect trials. Misses take place when S2 stimuli are categorized as coming from the S1 distribution; false alarms occur when S1 stimuli are judged to come from S2.

3.3. Confidence

Perceiving or, more specifically, classifying perceptual evidence *p* as S1 or S2 always involves some degree of uncertainty. Confidence in one's perception can also be characterized as resulting from a criterion-setting process (fig. 2, dashed lines). Confidence in a perceptual decision is determined by setting confidence criteria that further partition the decision space. When the perceptual response crosses both the detection criterion and the confidence criterion, observers report detecting the stimulus with high confidence (fig. 2A). If the perceptual response crosses the detection criterion but fails to cross the confidence criterion, observers correctly report detecting the target but with low confidence (fig. 2C). An analogous explanation in the other direction applies too. Observers report not detecting the stimulus with low confidence if the perceptual response crosses the left confidence criterion, but not the detection criterion (fig. 2B). When the internal response is too weak to cross any criteria, observers judge with high confidence that the stimulus is absent.[10]

[10] Whether this uncertainty is reflected in subjects' phenomenology or not is a matter of current debate. Recently, the question of whether there is perceptual confidence, or more generally whether perceptual experiences reflect the probabilistic nature of perception, has been widely discussed (Morrison [2016], [2017]; Munton [2016]; Block [2018]; Beck [2019]; Gross [2020]; Siegel [2022]). Here, I am neutral as to whether perceptual phenomenology reflects perceptual confidence or not. Confidence judgements could be based on perceptual confidence, but the SDT apparatus does not require this to be the case. For discussion on the relation between confidence and consciousness, see (Morales and Lau [2022]).

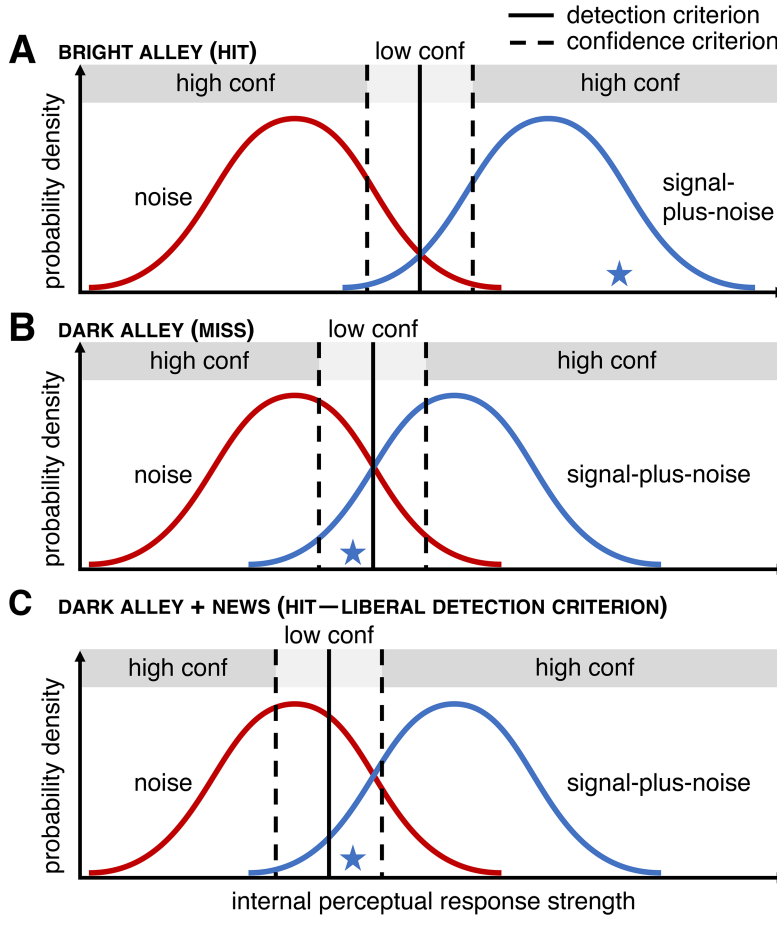

Figure 2. SDT models of alley scenarios. The distance between the noise and the signal-plus-noise curves represents the observer's perceptual sensitivity (*d'*) (for distributions with the same variance). (A) Bright alley scenario: The internal perceptual response (star) produced by the stimulus (the man) in a given trial crosses both the detection criterion and the right confidence criterion producing an accurate judgement with high confidence. (B) Dark alley scenario: The observer inaccurately judges the alley as being empty, albeit with low confidence because the perceptual response falls between the left-most confidence criterion and the detection criterion. (C) Dark alley plus news scenario: An identical perceptual response with an identical sensitivity as in the dark alley scenario produces an accurate detection (still with low confidence) due to left-shifted more liberal criteria.

This brief introduction to SDT highlights the crucial role the internal perceptual response plays in modelling perceptual judgements. To successfully explain introspection using insights from SDT, iSDT needs an equivalent notion: an internal introspective response. In the next section, I offer a plausible candidate for a basis

of introspective responses: conscious experience intensity or what I call 'mental strength' (Morales [2023]).

4. Mental Strength and Introspective Internal Responses

The targets of perception (stimuli) have degrees of strength: the face of a man in an alley can be more or less bright, sounds can be more or less loud, heat patches can be more or less hot, and so on. Stimuli in each modality may be strong (or weak) along more than one dimension; for example, visual stimuli strength depends on brightness, contrast, saturation, and so on. SDT postulates that after hitting our senses, stimuli produce an internal perceptual response of, *ceteris paribus*, proportional strength. In other words, strong stimuli typically produce strong perceptual responses, and weak stimuli typically produce weak perceptual responses. As explained in section 3, the probability of making an accurate perceptual judgement largely correlates, *ceteris paribus*, with the strength of the internal perceptual response (fig. 3A).

To model introspective sensitivity the way SDT models perceptual sensitivity, iSDT needs a functional analogue of internal perceptual responses. These are postulates of SDT (hidden variables) to explain perception, and hence iSDT can also postulate an internal introspective response that plays an analogous role when modelling introspective sensitivity. But what, if anything, produces introspective responses?

According to iSDT, the intensity of conscious experiences—their mental strength—modulates the strength of introspective responses, which, in turn, modulate the accuracy of introspective judgements. Conscious experiences vary in their degree of intensity. Pains can be stronger or weaker, mental images can be more or less vivid, perceptual experiences can be more or less intense (fig. 3B). iSDT relies on this obvious fact about our conscious experiences to calibrate introspective reliability.[11]

Mental strength is the phenomenal magnitude of conscious experiences. As such, the degree of strength of a conscious experience is its degree of phenomenal intensity. It increases from zero, as it were, when the conscious experience has not yet arisen, and grows in certain time to a given measure. Different degrees of mental strength result in different degrees with which mental events make their way to our consciousness. To express these same ideas slightly differently, mental strength is the degree of prominence that a conscious experience has in one's phenomenal field at a given time.[12]

[11] How do we know that conscious experiences have degrees of intensity? One may worry that if we know this introspectively, then any explanation of introspective accuracy based on mental strength may be compromised. I do not think we need to worry about this. You may fail to accurately perceive the exact brightness (and other properties) of a series of stimuli and yet accurately (and confidently) perceive that they differ along the brightness dimension. Inaccurate perception does not preclude us from perceiving that a series of stimuli differ among themselves in the misperceived dimension. Similarly, we could fail to introspect the intensity (and other properties) of our experiences and yet accurately (and confidently) introspect that they have different intensities. It is just this relatively uncontroversial fact about the degrees of intensity of conscious experiences that iSDT relies on.

[12] See Hill's ([1988]) 'volume control hypothesis' for a similar description of mental strength as variation in the prominence of a conscious experience in the phenomenal field.

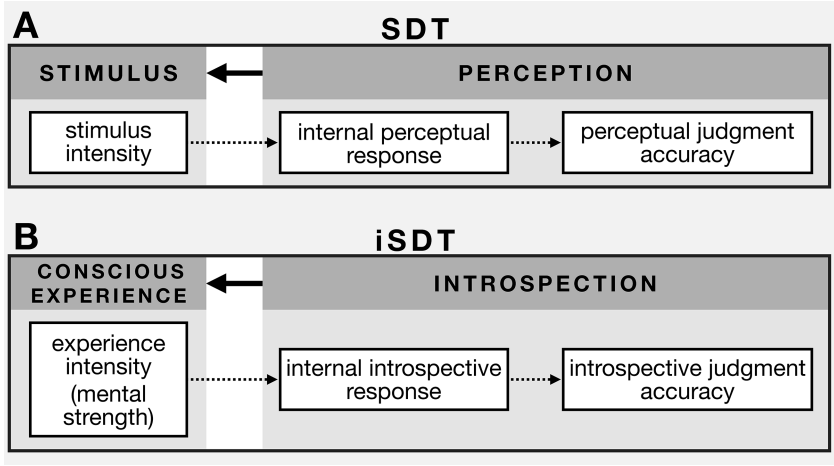

Figure 3. Internal responses as modelled by SDT and iSDT. (A) SDT. When perceiving a stimulus, its intensity largely modulates (but does not determine) the strength of the internal perceptual response in the perceiver. Perceptual accuracy is largely driven by the strength of the internal perceptual response (perceptual signal-to-noise ratio). (B) iSDT. iSDT offers an analogous explanation. When introspecting a conscious experience, its intensity or mental strength largely modulates (but does not determine) the strength of the internal introspective response in the introspector. Introspective accuracy is largely driven by the strength of the internal introspective response (introspective signal-to-noise ratio). Dotted arrows indicate non-deterministic modulation. Solid thick arrows indicate the target of each capacity.

Thus, an intense pain 'takes over' a larger portion of one's phenomenal field than a mild pain; a vivid mental image has more mental strength than a faint one; an experience of a loud sound has more mental strength than an experience of a quiet sound.[13]

Strong stimuli normally produce experiences with a strong internal response and, in turn, with strong mental strength (and vice versa for weak stimuli) (Peters and Lau [2015]).[14] Under normal circumstances, the larger a (potential) tissue damage is, the stronger the pain. The same applies for perception: typically, the stronger the stimulus, the stronger the perceptual experience. In visual imagery there is no external stimulation, but the clearer, sharper, more detailed and vivid the imagined objects

[13] Kant seems to hold a similar view regarding conscious intensive magnitudes in both the *Critique of Pure Reason* ('The anticipation of perception') and in his *Lectures on Metaphysics*: 'For example, when a representation has inhibited many others, we say that this has made a great impression' (cited in Longuenesse [1998], p. 320). Longuenesse's ([1998], p. 320) commentary of this passage is illuminating: 'Even states of consciousness can thus be [. . .] compared as to their magnitude. A representation is "more or less intense" according to the multiplicity of representations it inhibits; a very great pain makes one deaf and blind toward any other representation'.

[14] What exactly makes internal perceptual responses conscious is a matter of contention in the philosophy and science of consciousness. This is not the place to take a stance in that debate since all we need is that there is a rough correlation in normal cases between stronger internal perceptual responses and more intense conscious experiences. Even though different views disagree about the exact conditions that make this correlation possible or the conditions under which it breaks down, most views would admit that a stronger internal perceptual response results in a more intense experience.

are, the more intense the visual image tends to be overall (for discussion of the nature of mental imagery vividness, see Kind [2017]; for a dissenting view on the nature of experience intensity in general, see Fazekas [2024]).

This correlation, however, does not always hold. Ever-present noise, the subject's overall internal state, deployed attention, familiarity with the stimuli, and other circumstances can weaken or even reverse this correlation. For example, there is no tissue to be (potentially) damaged in a missing limb, and yet phantom limb pains can be intense. Conversely, when adrenaline is really high, large tissue damage may produce little to no pain. Similarly, a vivid mental image of a very faint candle flame in a dark room lacks clear details by necessity (otherwise it would not be a mental image with those contents).

The intensity of perceptual experiences can be similarly decoupled from the stimulus intensity that generates them. Extreme silence produces intense auditory experiences (Sorensen [2009]; Cox [2014]). Something similar happens in the visual domain during 'subjective inflation' (Knotts et al. [2019]).[15] Peripheral vision is not as rich as foveal vision, which results in weaker internal responses pertaining to peripheral stimuli. Subjects may nonetheless enjoy intense and detailed experiences in the periphery—sometimes even more than in foveal regions (Odegaard et al. [2018]). Subjective inflation can even make subjects fail to notice drastic changes that impoverish stimuli in the periphery (Cohen et al. [2020]). Naturally, these inflated, more intense experiences do not reflect the true nature of the stimulus. Peripheral vision rarely feels drastically impoverished: it is not experienced in black and white or with dramatically washed-out colours, and people are normally confident—in fact, overconfident—about their discrimination capacities in peripheral vision. Nevertheless, peripheral perception is in fact drastically impoverished (for example, colour discrimination is poor). This is a clear case where the internal perceptual response is weak and yet mental strength is strong. Alternatively, blindsight patients display highly accurate unconscious perception (which requires strong perceptual responses) that, however, does not lead to a conscious experience (and, thereby, mental strength is lacking altogether) (Weiskrantz [1986]).[16]

Relatedly, internal responses with identical signal-to-noise ratios may create experiences with different degrees of mental strength. In a recent experimental paradigm called 'matched-performance/different-confidence', specifically designed stimulus pairs yield matched performance in an experimental task while producing significant differences in subjects' confidence ratings in their performance (Koizumi et al. [2015]; Samaha et al. [2016]; Maniscalco et al. [unpublished]). Matched performance is achieved by matching the signal-to-noise ratio of two stimuli that, nevertheless, differ in their overall energy. These stimuli generate matched internal perceptual responses, thus making it equally difficult to discriminate the signal. But, at the same time, the

[15] For a recent debate, see (Abid [2019]; Knotts et al. [2020]).

[16] Note that this is true even if blindsight is reinterpreted as qualitatively degraded conscious vision (Phillips [2021]).

stimulus with more overall energy looks more intense (for example, the contrast looks more marked).[17] Likely, this increase in mental strength is part of what makes subjects rate their otherwise identical performance with higher confidence.

Following these perceptual scenarios, iSDT postulates that internal introspective responses mostly (but not solely) are modulated by the strength of conscious experiences. In normal cases, intense experiences produce strong introspective responses. But due to noise, a weak experience could occasionally generate a strong introspective response, or a strong experience could occasionally generate a weak introspective response. Following the perceptual case, iSDT stipulates that there is a close modulation of introspective responses by mental strength, but not a perfect correlation. Moreover, since mental strength does not always depend on stimulus intensity, the strength of introspective responses does not always depend on stimulus intensity either (see the next section for examples).

The details of a theory of mental strength need not be further specified here (instead, see Morales [2023]). All we need to sketch a model of introspective accuracy is the notion of an introspective response that is modulated by the intensity of conscious experiences, which I have provided here.

5. Introspective Signal Detection Theory

5.1. A case study: Pain

We now have the necessary building blocks to present iSDT and how it models introspective sensitivity (as well as response bias and confidence) in a systematic manner. I start with pain introspection as a case study. In the next section, I expand the framework to introspection of visual imagery and perceptual experiences.

Consider the following scenarios, all of which assume you are in fact experiencing pain:[18]

Strong Pain: You wake up with a very strong toothache. You rush to the dentist. They ask if you are sure you are in pain. You introspect your experience and accurately judge that you are indeed experiencing a strong dental pain. You are highly confident. The dentist's question even seems odd—of course you are confident you are in strong pain!

Mild Pain: An hour after taking powerful painkillers, your toothache becomes quite mild. When you introspect, you honestly—albeit inaccurately—judge that you are not in pain anymore. When the dentist asks if you are sure, the question does not seem as odd as before: you are legitimately not completely sure (or, in any case, your confidence is lower than in the strong pain scenario).

[17] For an example of this kind of stimuli, see (Samaha et al. [2016], fig. 1A).

[18] For simplicity, in what follows I discuss pain detection (whether you are or not in pain), but the iSDT machinery can be equally applied to discrimination (for example, whether a pain is throbbing or stabbing).

Mild Pain + Familiarity: This scenario is identical to the mild pain scenario (by stipulation, the intensity of the pain is identical in both scenarios). Here, however, you are familiar with dental pains. And because you have experienced them before, you know it usually takes several hours for the painkillers to make them go completely away. This time you accurately introspect that you are still in pain, but you are not very confident.

Now consider the following scenario, which assumes that you are not experiencing any pain:

Dental Fear: You go to the dentist for a routine cleanup. You cannot experience any pain because you are under a powerful local anaesthetic. But you have always been really scared of these procedures. The dentist turns on their loud and scary instruments, and as they approach your mouth, you start to closely monitor your experience. Suddenly, you yelp and report feeling an intense pain. The dentist is confused: not only are you under the anaesthetic, but also they have not even touched you yet.[19]

The strong pain scenario—and to some extent the mild pain + familiarity scenario—seem quite plausible, but, admittedly, the mild pain and dental fear scenarios may strike some as counterintuitive. How could you be in pain and miss it? How could you not be in pain and think you are? The intuitiveness of these and other cases, however, cannot be assessed introspectively under risk of getting them wrong (the garden-variety assumption makes this an open possibility). Introspective inaccuracies are normally not accessible through introspection, and thus it might never seem to oneself that an introspective mistake is taking place. Introspective mistakes are also not (easily) corrigible by others (Rorty [1970]; Alston [1971]; Dennett [2002]; Langland-Hassan [2017]), unlike perceptual errors that can easily be pointed out by someone else. Thus, intuition, introspection, or the lack of correction from others are not good routes to discover introspective errors or whether they are possible. In contrast, a wide-scope and illuminating theory should allow us to reason through these scenarios and establish how they work in a principled manner. The purpose of introducing these scenarios—some quite normal, some *prima facie* far-fetched albeit perfectly consistent with the garden-variety assumption—is to show how iSDT can explain all of them in a systematic and principled way.

Similarly to how perceptual judgements are conceived in SDT, introspective judgements in iSDT are determined by sensitivity (the ability to discriminate experiences based on the way these shape a psychological decision space) and by response biases (the manner in which the psychological space is partitioned to generate possible responses). Introspecting is modelled as an introspector deciding whether an internal

[19] Not only anaesthetized patients experience dental fear. Patients whose tooth's nerves have been removed may experience it too (Rosenthal [2005], p. 127; Meier et al. [2014]). In these cases, experiencing actual physical pain—let alone intense pain—should be short from impossible, and an alternative explanation for the pain report—such as an introspective error—is needed.

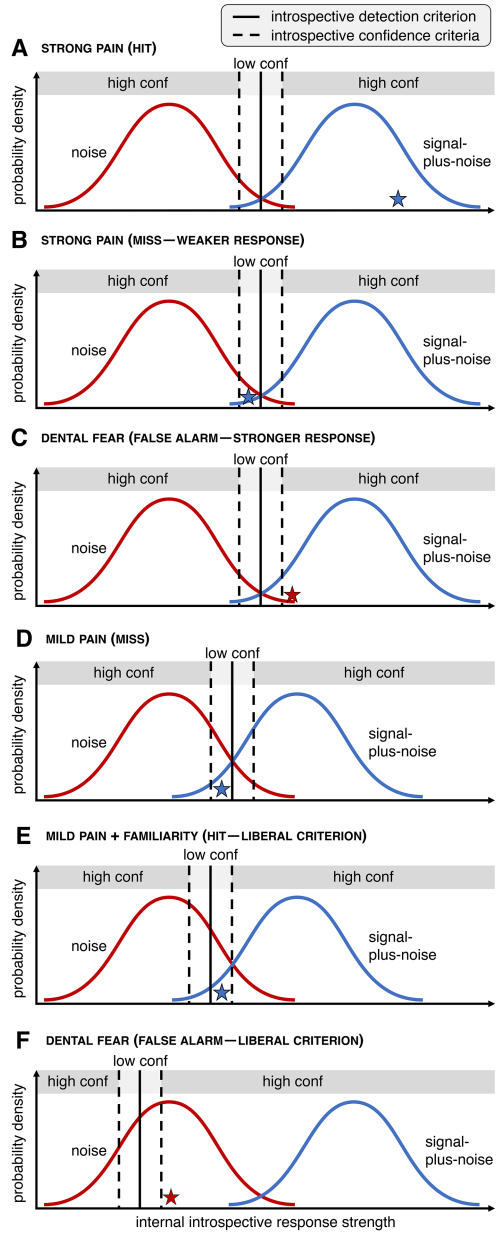

Figure 4. Pain scenarios as modelled by iSDT. (A) Strong pain scenario (hit): The barely overlapping distributions indicate high sensitivity; a strong introspective response *i* (dark grey star indicating *i* is drawn from the signal-to-noise distribution) is accurately and confidently classified as pain. (B) Strong pain scenario (miss): Random factors that weaken the introspective response during an identical strong pain result in inaccurate

introspective response *i* was generated by a conscious-experience class C1 (for example, 'pain absent', 'burning pain', and so on) or C2 (for example, 'pain present', 'stabbing pain', and so on). The introspective response corresponds to the strength of the introspective evidence, in turn modulated by the intensity of the conscious experience (its mental strength). Repeated experiences of the same class produce introspective responses with different values due to ever-present noise of different sorts. The values of the introspective response are distributed across a decision axis. The introspective response *i* in any given case can be thought of as being drawn from either a noise or a signal-plus-noise distribution (fig. 4).

5.2. Introspective sensitivity

The distance between the distributions' means determines the introspector's sensitivity.[20] In the strong pain scenario, the distributions do not overlap much, indicating—as expected—high introspective sensitivity for strong pains (fig. 4A). An intense dental pain produces a strong introspective response that is easy to introspectively judge as a strong pain (it is far from the detection criterion).

Note that the probability of error in the strong pain scenario is very small (the area of overlap between the two curves in fig. 4A). Introspective errors under these conditions should be quite rare, but not impossible (in the same way that it is rare to fail to see a man in a bright alley who is in your line of sight, but not impossible). Introspective misses of a strong pain, for example, are possible if the introspective response fails to cross the detection criterion (strong pain (miss) scenario; fig. 4B). This could happen even if the mental strength of a particular painful experience is strong. Recall

[20] With background assumptions about the type and variance of the distributions.

introspection albeit with low confidence. (C) Dental fear (false alarm—stronger response): Fear of a dental procedure increases the introspective response of a non-painful experience (light grey star indicating *i* is drawn from the noise distribution). In an extremely unlikely—but possible—scenario, the introspective response crosses both the detection and the confidence criteria resulting in a high-confidence introspective false alarm. (D) Mild pain (miss) scenario: The average mild pain has a weaker internal response, which is depicted here as a lower mean of the signal distribution that results in an increased overlap with the noise distribution. This indicates lower introspective sensitivity for this kind of pains. Here, a mild pain with introspective evidence *i* (dark grey star) is missed but with low confidence. (E) Mild pain plus familiarity scenario (hit—liberal criterion): Slightly more liberal criteria result in a different (this time accurate) classification of the introspective response. Despite the criterion shift, confidence is still low. (F) Dental fear (false alarm—liberal criterion): The introspective response of the lack of pain (noise) is prototypically weak but, due to fear, here the criteria are drastically shifted, making them more liberal. As in C, the introspector confidently misclassifies the introspective response as coming from the signal-plus-noise (pain) distribution.

that the relation between mental strength and introspective response is not deterministic (fig. 3B). The process that gives rise to an introspective response from a strong experience may get corrupted, generating a weak introspective response (that is, weaker than what that kind of pain normally generates). In a case like this, iSDT predicts that although you are experiencing a strong pain, you introspectively judge that you are not.

Introspective false alarms of strong pains are possible too (fig. 4C). The iSDT framework has a straightforward way of accommodating rare cases such as the dental fear scenario. Patients' fear of the procedure in conjunction with vibrations produced by the dentist's instruments (noise in iSDT terms) may significantly increase the introspective response. This, in turn, produces a pain report even though no pain is (could be) experienced under these circumstances (namely, under potent anaesthesia or without dental nerves). (See the next subsection for an explanation of the dental fear scenario that appeals to response bias instead of increased introspective response.)

The situation for mild pains is quite different. Mild pains' mental strength tends to be less intense, which in turn tends to make introspective responses weaker. In consequence, introspective sensitivity in the mild pain scenario is lower. This can be modelled by decreasing the mean of the signal-plus-noise distribution, making the distributions overlap more (fig. 4D). This does not necessarily entail that false alarms and misses are frequent, just that we should expect them to be less rare than during introspection of strong pains.

5.3. Introspective response bias

Another advantage of iSDT is that we can keep introspective sensitivity and response bias apart. A full calibration of introspection's range of reliability requires us to consider response biases. Accurate and inaccurate introspective judgements may arise not (only) because of an insensitive or inaccurate machinery, but because of a suboptimal decision rule to classify the relevant introspective signal.[21]

Criterion effects can explain introspective variation, even when holding introspective sensitivity fixed. In the mild pain and mild pain + familiarity scenarios, introspective sensitivity and internal responses are, by stipulation, identical (figs. 4D and 4E). And yet, in the former scenario the pain is not introspectively detected, while it is in the latter. iSDT models these scenarios by shifting the introspective criterion. In the mild pain + familiarity scenario, knowledge about the time course of the painkillers makes your introspective criterion more liberal; you know it is unlikely to be free of pain so quickly, so you are willing to judge being in pain with less introspective evidence. In contrast, in the mild pain scenario you have an unbiased detection criterion, and the same weak introspective response is insufficient for introspecting

[21] Perceptual judgement mistakes can often be explained by suboptimal response biases too (Rahnev and Denison [2018]).

that you are in pain. Finally, the dental fear scenario could also be explained by a liberal shift in the criterion (fig. 4F), instead of explaining it as an increased introspective noise response, as suggested above (fig. 4C). According to this explanation, the fearful patient is more willing to classify as pain a really weak introspective response produced by vibrations (introspective noise). Naturally, a combination of an increased introspective response and a more liberal criterion is possible too.

In its current form, iSDT aims to leverage SDT's insights rather than its strict mathematical formulations. A huge advantage of SDT for measuring sensitivity is that d' incorporates both hits and false alarms rather than raw percentage correct (sec. 3). iSDT can take this insight to refine the way we think about introspective sensitivity and response biases beyond a raw accurate/inaccurate classification of introspective judgements. One consequence of having a liberal criterion in the mild pain + familiarity scenario is not only that existing mild pains are more easily detected (higher hit rate), but also that it is implied that more false alarms (less correct rejections) are possible. In other words, someone with a liberal introspective criterion might detect more (or even all) their relevant pain experiences, but they would do so at the cost of increasing their false alarms. The opposite is true for someone with a conservative criterion: they might rarely (or never) raise a false alarm, but they would do so at the cost of increasing the number of times they miss some experiences they try to introspect. This is a notable consequence of iSDT that can help model introspective behaviour in quite a subtle way (perhaps subtler than some current philosophical approaches to introspection allow), at the same time that it makes introspection's machinery consistent with that of other faculties.

5.4. Introspective confidence

Infallibilists and sceptics alike have taken excess and lack of confidence, respectively, as evidence to support their views. In contrast, iSDT can explain these variations in introspective confidence in a way that is largely orthogonal to the reliability of introspection.[22] As in SDT, confidence in an introspective judgement is a function of the strength of the introspective response and the placement of confidence criteria (dashed lines in fig. 4).[23] To capture the common intuition that introspection is, if not infallible, unlikely to be significantly wrong most of the time (*pace* Schwitzgebel and other

[22] Some of these confidence variations are reflected in how we talk. For example, subjects use 'I feel pain' more often to describe minor pains, and 'I have/am in pain' to describe major pains (Reuter [2011]).

[23] Snodgrass and Shevrin ([2006]) describe a similar situation; their focus, however, is on perceptual responses, not on what I call here introspective responses. According to them, trials whose internal perceptual response falls between a detection criterion and a subjective criterion (akin to a confidence criterion) can be labelled as 'weak consciousness trials'. These are phenomenally conscious trials that are, however, not access-conscious (Block [1995], [2007]). On Snodgrass and Shevrin's view, this would entail that they are not amenable to introspection, which requires the orientation of attention and forming a judgement. My view is different from theirs because I think that trials that give rise to experiences with little mental strength, and therefore with a small introspective response, can still be introspected, albeit more inaccurately. For a criticism of Snodgrass and Shevrin's view, see (Irvine [2009]).

sceptics), confidence criteria in figure 4 are placed much closer to the detection criterion than they were in figure 2 for the perceptual cases. This entails introspective responses stay in the high-confidence regions of the introspective decision axis in most cases.

Importantly, at least sometimes, introspective judgements are made with low confidence. This fact has been used by sceptics to mount their generalized attacks on introspective reliability (the mild pain and mild pain + familiarity scenarios [figs. 4D and 4E]). Low confidence, however, is not necessarily the best guide to establish the reliability of a detection system (an obvious fact from the separability of sensitivity and response bias in SDT and iSDT). As the scenarios above show, (accurate) introspective judgements may still be made with low confidence even when introspection is highly reliable. Of course, the opposite is true as well; in cases of lower sensitivity, subjects may introspect with high confidence.

The analysis of these scenarios shows how iSDT satisfies the two desiderata described at the outset. It explains accurate and inaccurate introspection under a wide range of circumstances. It also explains why this is the case in a systematic and illuminating way by appealing to a single kind of explanation that can accommodate otherwise disparate cases. Moreover, iSDT can also model (and predict!) important features that drive our introspective behaviour such as response biases and confidence. Finally, the iSDT framework is also a reminder of the importance of letting theory—not intuition or introspection—lead the way in how we reason through complex and, at least *prima facie*, unintuitive scenarios.[24]

6. Beyond Pains: Mental Imagery and Perception

iSDT models introspection in the same way for any conscious experience with a degree of intensity (mental strength) that generates an introspective response.[25] Here, I can only briefly sketch how to expand iSDT to mental imagery and visual perception. Rather than a full treatment of how to apply iSDT to these cases, this sketch is meant as a proof of concept that the machinery developed in the previous section is helpful for constraining and guiding our thinking about the scope of reliability of introspection of other kinds of conscious experience.

A particularly vivid mental image of a simple object (for example, a red apple) is an instance of an intense experience in the imagery domain. iSDT predicts cases like this produce strong introspective responses that result in accurate, confident introspective judgements. But conjuring vivid mental images is hard. When these are faint, such that attending to their features becomes hard, introspecting them may become harder too. More inaccurate judgements with lower confidence can be expected. It is a matter of contention what makes an image more or less vivid, or even what is

[24] For a recent theoretical evaluation of iSDT, see (Dołęga [2023]).

[25] Other experiences that could be modelled by iSDT include itches, emotions, moods, action-awareness, sense of bodily ownership, and—if they have distinctive phenomenal consciousness—perhaps thoughts, desires, memories, and mental effort (Doulatova [2019]).

meant by vividness (Kind [2017]; Fazekas [2023]). However, it is reasonable to assume that overall mental strength of a mental image is the result of an intensity aggregate across several dimensions: sensory properties (brightness, loudness, and so on); clarity, number, and salience of details; the feeling of presence of the imagined objects or events; and the overall stability of the image (Cornoldi et al. [1991]).

Perceptual experiences can be expected to follow exactly the same pattern. As discussed above, on one hand, there is a strong link between perceptual response and mental strength and, on the other hand, between mental strength and introspective response. These links predict that we are likely to introspect accurately strong experiences of strong stimuli. But these links can be broken. In principle, even strong perceptual experiences can be inaccurately introspected based on serendipitous weak introspective responses. Moreover, subjective inflation (Odegaard et al. [2018]) or Sperling and Sperling-like phenomena (Sperling [1960]; Landman et al. [2003]; Sligte et al. [2008]) suggest weak perceptual responses can produce experiences with high mental strength. iSDT predicts that in these cases, participants are likely to introspect accurately their strong experiences even when their introspective reports do not reflect the weakened nature of external stimuli and internal perceptual processing. Thinking about these cases in this way allows us to reinterpret phenomenal overflow, according to which phenomenology exceeds the capacity of cognitive access (Block [1995], [2007]). On this reinterpretation (Knotts et al. [2019]), subjects do not fail to access their phenomenal contents, and they do not introspect their experiences inaccurately either. Rather, subjects accurately introspect rich (inflated) experiences that are, nevertheless, not supported by perceptual processing. It is poor perceptual processing that explains subjects' inability to report stimuli accurately, not lack of cognitive access.

7. Conclusion

Introspection is signal detection. iSDT explains why, sometimes, we introspect accurately; and it also explains why, sometimes, we can expect introspection to be inaccurate. In doing so, iSDT validates some of the intuitions of extreme, incompatible views that hold introspection is infallible or utterly unreliable. I take this to be a virtue of the proposal. A huge advantage of iSDT is that it offers a detailed, systematic, naturalistic, and psychologically plausible explanation of introspection's whole range of reliability. Importantly, it achieves this in an illuminating way—it explains why accurate and inaccurate cases take place—and it does so in an elegant way appealing to a single mechanism. This introspective machinery operates, at a fundamental level, in similar ways to other faculties, such as perception or memory, which have been successfully modelled in psychology. By comparing perceptual stimulus strength to mental strength, I showed that the tools developed by signal detection theory provide a novel and solid theoretical scaffolding for modelling variations in introspective sensitivity, response bias, and confidence.

Acknowledgements

Thanks to David Barack, Simon Brown, Chaz Firestone, Nemira Gasiunas, Matthew Heeney, Hawkan Lau, John Morrison, Elliot Paul, Christopher Peacocke, Ian Phillips, and Wayne Wu, as well as two anonymous reviewers, for helpful comments on earlier versions of the article, and to audiences at Columbia University, Johns Hopkins University, the 2017 Society for Philosophy and Psychology, and the 2019 Association for the Scientific Study of Consciousness for their feedback.

Department of Psychology
and
Department of Philosophy
Northeastern University
Boston, MA, USA
j.morales@northeastern.edu

References

Abid, G. [2019]: 'Deflating Inflation: The Connection (or Lack Thereof) between Decisional and Metacognitive Processes and Visual Phenomenology', *Neuroscience of Consciousness*, **5**, available at <doi.org/10.1093/nc/niz015>.

Alston, W. [1971]: 'Varieties of Privileged Access', *American Philosophical Quarterly*, **8**, pp. 223–41.

Armstrong, D. M. [1968]: *A Materialist Theory of the Mind*, London: Routledge and Kegan Paul.

Bayne, T. and Spener, M. [2010]: 'Introspective Humility', *Philosophical Issues*, **20**, pp. 1–22.

Beck, J. [2019]: 'On Perceptual Confidence and "Completely Trusting Your Experience"', *Analytic Philosophy*, **59**, pp. 385–15.

Block, N. [1995]: 'On a Confusion about a Function of Consciousness', *Behavioral and Brain Sciences*, **18**, pp. 227–47.

Block, N. [2007]: 'Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience', *Behavioral and Brain Sciences*, **30**, pp. 481–548.

Block, N. [2018]: 'If Perception Is Probabilistic, Why Does It Not Seem Probabilistic?', *Philosophical Transactions of the Royal Society B*, **373**, available at <doi.org/10.1098/rstb.2017.0341>.

Carrasco, M. [2011]: 'Visual Attention: The Past 25 Years', *Vision Research*, **51**, pp. 1484–525.

Carrasco, M., Ling, S. and Read, S. [2004]: 'Attention Alters Appearance', *Nature Neuroscience*, **7**, pp. 308–13.

Carruthers, P. [2000]: *Phenomenal Consciousness*, Cambridge: Cambridge University Press.

Chalmers, D. [2003]: 'The Content and Epistemology of Phenomenal Belief', in Q. Smith and A. Jokic (*eds*), *Consciousness: New Philosophical Perspectives*, New York: Oxford University Press, pp. 220–72.

Chalmers, D. J. [2010]: *The Character of Consciousness*, New York: Oxford University Press.

Chirimuuta, M. [2014]: 'Psychophysical Methods and the Evasion of Introspection', *Philosophy of Science*, **81**, pp. 914–26.

Cohen, M. A., Botch, T. L. and Robertson, C. E. [2020]: 'The Limits of Color Awareness during Active, Real-World Vision', *Proceedings of the National Academy of Sciences USA*, **117**, pp. 13821–27.

Cornoldi, C., De Beni, R., Giusberti, F. and Marucci, E. [1991]: 'The Study of Vividness of Images', in R. H. Logie (*ed*.), *Mental Images in Human Cognition*, Amsterdam: North Holland.

Cortese, A., Amano, K., Koizumi, A., Kawato, M. and Lau, H. [2016]: 'Multivoxel Neurofeedback Selectively Modulates Confidence without Changing Perceptual Performance', *Nature Communications*, **7**, available at <doi.org/10.1038/ncomms13669>.

Cox, T. [2014]: *The Sound Book*, New York: W. W. Norton.

Dennett, D. C. [2002]: 'How Could I Be Wrong? How Wrong Could I Be?', *Journal of Consciousness Studies*, **5–6**, pp. 13–16.

Descartes, R. [1984]: *The Philosophical Writings of Descartes*, Cambridge: Cambridge University Press.

Dołęga, K. [2023]: 'Models of Introspection vs. Introspective Devices Testing the Research Programme for Possible Forms of Introspection', *Journal of Consciousness Studies*, **30**, pp. 86–101.

Doulatova, M. [2019]: 'Tracking Intentionalism and the Phenomenology of Mental Effort', *Synthese*, **115**, pp. 330–17.

Estes, W. K. [2002]: 'Traps on the Route to Models of Memory and Decision', *Psychonomic Bulletin and Review*, **9**, pp. 3–25.

Fazekas, P. [2024]: 'Vividness and Content', *Mind and Language*, **39**, pp. 61–79.

Fleming, S. M., Ryu, J., Golfinos, J. G. and Blackmon, K. E. [2014]: 'Domain-Specific Impairment in Metacognitive Accuracy following Anterior Prefrontal Lesions', *Brain*, **137**, pp. 2811–22.

Gertler, B. [2001]: 'Introspecting Phenomenal States', *Philosophy and Phenomenological Research*, **63**, pp. 305–28.

Giustina, A. and Kriegel, U. [2017]: 'Fact-Introspection, Thing-Introspection, and Inner Awareness', *Review of Philosophy and Psychology*, **8**, pp. 143–64.

Goldman, A. [2004]: 'Epistemology and the Evidential Status of Introspective Reports', *Journal of Consciousness Studies*, **11**, pp. 1–16.

Goldman, A. [2006]: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, New York: Oxford University Press.

Gross, S. [2020]: 'Probabilistic Representations in Perception: Are There Any, and What Would They Be?', *Mind and Language*, **35**, pp. 377–89.

Hatfield, G. [2005]: 'Introspective Evidence in Psychology', in P. Achinstein (*ed*.), *Scientific Evidence: Philosophical Theories and Applications*, Baltimore, MD: Johns Hopkins University Press, pp. 259–86.

Hill, C. S. [1988]: 'Introspective Awareness of Sensations', *Topoi*, **7**, pp. 11–24.

Hohwy, J. [2011]: 'Phenomenal Variability and Introspective Reliability', *Mind and Language*, **26**, pp. 261–86.

Horgan, T. and Kriegel, U. [2007]: 'Phenomenal Epistemology: What Is Consciousness That We May Know It So Well?', *Philosophical Issues*, **17**, pp. 123–44.

Humphreys, P. [2002]: 'Computational Models', *Philosophy of Science*, **69**, pp. S1–11.

Irvine, E. [2009]: 'Signal Detection Theory, the Exclusion Failure Paradigm, and Weak Consciousness: Evidence for the Access/Phenomenal Distinction?', *Consciousness and Cognition*, **18**, pp. 551–60.

Kant, I. [1998]: *Critique of Pure Reason*, Cambridge: Cambridge University Press.

Kind, A. [2017]: 'Imaginative Vividness', *Journal of the American Philosophical Association*, **3**, pp. 32–50.

Knotts, J. D., Michel, M. and Odegaard, B. [2020]: 'Defending Subjective Inflation: An Inference to the Best Explanation', *Neuroscience of Consciousness*, **2020**, available at <doi.org/10.1093/nc/niaa025>.

Knotts, J. D., Odegaard, B., Lau, H. and Rosenthal, D. [2019]: 'Subjective Inflation: Phenomenology's Get-Rich-Quick Scheme', *Current Opinion in Psychology*, **29**, pp. 49–55.

Knuuttila, T. and Loettgers, A. [2014]: 'Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines', *The Monist*, **97**, pp. 280–300.

Knuuttila, T. and Loettgers, A. [2016]: 'Model Templates within and between Disciplines: From Magnets to Gases—and Socio-Economic Systems', *European Journal for Philosophy of Science*, **6**, pp. 377–400.

Koizumi, A., Maniscalco, B. and Lau, H. [2015]: 'Does Perceptual Confidence Facilitate Cognitive Control?', *Attention, Perception, and Psychophysics*, **77**, pp. 1295–306.

Landman, R., Spekreijse, H. and Lamme, V. A. F. [2003]: 'Large Capacity Storage of Integrated Objects before Change Blindness', *Vision Research*, **43**, pp. 149–64.

Langland-Hassan, P. [2017]: 'Pain and Incorrigibility', in J. Corns (*ed.*), *The Routledge Handbook of Philosophy of Pain*, London: Routledge.

Lin, C.-H. [unpublished]: 'Tool Migration: A Framework for Analyzing Cross-Disciplinary Use of Mathematical Constructs', available at <philsci-archive.pitt.edu/14929/>.

Locke, J. [1975]: *An Essay Concerning Human Understanding*, Oxford: Clarendon.

Longuenesse, B. [1998]: *Kant and the Capacity to Judge*, Princeton, NJ: Princeton University Press.

Lycan, W. G. [1996]: *Consciousness and Experience*, Cambridge, MA: MIT Press.

Macmillan, N. A. and Creelman, C. D. [2005]: *Detection Theory*, Mahwah, NJ: Lawrence Erlbaum.

Maniscalco, B., Graham Castaneda, O., Odegaard, B., Morales, J., Rajananda, S. and Peters, M.A.K. [unpublished]: 'The Metaperceptual Function: Exploring Dissociations between Confidence and Task Performance with Type 2 Psychometric Curves', available at <psyarxiv.com/5qrjn/>.

Maxwell, J. C. [1861]: 'On Physical Lines of Force, Part I: The Theory of Molecular Vortices Applied to Magnetic Phenomena', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **21**, pp. 161–75.

Meier, M. L., De Matos, N. M. P., Brügger, M., Ettlin, D. A., Lukic, N., Cheetham, M., Jäncke, L. and Lutz, K. [2014]: 'Equal Pain—Unequal Fear Response: Enhanced Susceptibility of Tooth Pain to Fear Conditioning', *Frontiers in Human Neuroscience*, **8**, available at <doi.org/10.3389/fnhum.2014.00526>.

Merton, R. C. [1969]: 'Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case', *Review of Economics and Statistics*, **51**, pp. 247–57.

Morales, J. [2023]: 'Mental Strength: A Theory of Experience Intensity', *Philosophical Perspectives*, **37**, pp. 248–68.

Morales, J. and Lau, H. [2022]: 'Confidence Tracks Consciousness', in J. Weisberg (*ed.*), *Qualitative Consciousness: Themes from the Philosophy of David Rosenthal*, Cambridge: Cambridge University Press, pp. 91–105.

Morales, J., Odegaard, B. and Maniscalco, B. [2022]: 'The Neural Substrates of Conscious Perception without Performance Confounds', in F. De Brigard and W. Sinnott-Armstrong (*eds*), *Neuroscience and Philosophy*, Cambridge, MA: MIT Press, pp. 285–323.

Morrison, J. [2016]: 'Perceptual Confidence', *Analytic Philosophy*, **57**, pp. 15–48.

Morrison, J. [2017]: 'Perceptual Confidence and Categorization', *Analytic Philosophy*, **58**, pp. 71–85.

Munton, J. [2016]: 'Visual Confidences and Direct Perceptual Justification', *Philosophical Topics*, **44**, pp. 301–26.

Odegaard, B., Chang, M. Y., Lau, H. and Cheung, S.-H. [2018]: 'Inflation versus Filling-In: Why We Feel We See More Than We Actually Do in Peripheral Vision', *Philosophical Transactions of the Royal Society B*, **373**, available at <doi.org/10.1098/rstb.2017.0345>.

Peacocke, C. [1998]: 'Nonconceptual Content Defended', *Philosophy and Phenomenological Research*, **58**, pp. 381–88.

Peters, M. A. K. and Lau, H. [2015]: 'Human Observers Have Optimal Introspective Access to Perceptual Processes Even for Visually Masked Stimuli', *eLife*, **4**, available at <https://doi.org/10.7554/eLife.09651.001>.

Phillips, I. [2021]: 'Blindsight Is Qualitatively Degraded Conscious Vision', *Psychological Review*, **128**, pp. 558–84.

Rahnev, D. and Denison, R. N. [2018]: 'Suboptimality in Perceptual Decision Making', *Behavioral and Brain Sciences*, **41**, available at <doi.org/10.1017/S0140525X18000936>.

Renero, A. [2019]: 'Modes of Introspective Access: A Pluralist Approach', *Philosophia*, **47**, pp. 823–44.

Reuter, K. [2011]: 'Distinguishing the Appearance from the Reality of Pain', *Journal of Consciousness Studies*, **18**, pp. 94–109.

Rorty, R. [1970]: 'Incorrigibility as the Mark of the Mental', *Journal of Philosophy*, **67**, pp. 399–424.

Rosenthal, D. [2005]: *Consciousness and Mind*, New York: Oxford University Press.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. and Lau, H. [2010]: 'Theta-burst Transcranial Magnetic Stimulation to the Prefrontal Cortex Impairs Metacognitive Visual Awareness', *Cognitive Neuroscience*, **1**, pp. 165–75.

Ryle, G. [2009]: *The Concept of Mind*, New York: Routledge.

Samaha, J., Barrett, J. J., Sheldon, A. D., Larocque, J. J. and Postle, B. R. [2016]: 'Dissociating Perceptual Confidence from Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory', *Frontiers in Psychology*, **7**, available at <doi.org/10.3389/fpsyg.2016.00851>.

Samuelson, P. A. [1969]: 'Lifetime Portfolio Selection by Dynamic Stochastic Programming', *Review of Economics and Statistics*, **51**, pp. 239–46.

Schurgin, M. W., Wixted, J. T. and Brady, T. F. [2020]: 'Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength', *Nature Human Behaviour*, **4**, pp. 1156–72.

Schwitzgebel, E. [2008]: 'The Unreliability of Naive Introspection', *Philosophical Review*, **117**, pp. 245–73.

Schwitzgebel, E. [2012]: 'Introspection, What?', in D. Smithies and D. Stoljar (*eds*), *Introspection and Consciousness*, New York: Oxford University Press, pp. 29–48.

Shoemaker, S. S. [1996]: *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press.

Siegel, S. [2022]: 'How Can Perceptual Experiences Explain Uncertainty?', *Mind and Language*, **37**, pp. 134–58.

Sligte, I. G., Scholte, H. S. and Lamme, V. A. F. [2008]: 'Are There Multiple Visual Short-Term Memory Stores?', *PLOS One*, **3**, available at <doi.org/10.1371/journal.pone.0001699>.

Smith, J. M. and Price, G. R. [1973]: 'The Logic of Animal Conflict', *Nature*, **246**, pp. 15–18.

Snodgrass, M. and Shevrin, H. [2006]: 'Unconscious Inhibition and Facilitation at the Objective Detection Threshold: Replicable and Qualitatively Different Unconscious Perceptual Effects', *Cognition*, **101**, pp. 43–79.

Sorensen, R. [2009]: 'Hearing Silence: The Perception and Introspection of Absences', in M. Nudds and C. O'Callaghan (*eds*), *Sounds and Perception*, Oxford: Oxford University Press, pp. 126–45.

Spener, M. [2015]: 'Calibrating Introspection', *Philosophical Issues*, **25**, pp. 300–21.

Sperling, G. [1960]: 'The Information Available in Brief Visual Presentations', *Psychological Monographs: General and Applied*, **74**, pp. 1–29.

Srinivasan, A. [2015]: 'Are We Luminous?', *Philosophy and Phenomenological Research*, **90**, pp. 294–319.

Stoljar, D. [2022]: 'Armstrong's Just-So Story about Consciousness', in P. Anstey and D. Braddon-Mitchell (*eds*), *A Materialist Theory of the Mind: 50 Years On*, Oxford: Oxford University Press, pp. 176–94.

Tanner, W. P., Jr [1954]: 'A Decision-Making Theory of Visual Detection', *Psychological Review*, **61**, pp. 401–9.

Weiskrantz, L. [1986]: *Blindsight: A Case Study and Implications*, Oxford: Oxford University Press.

Williamson, T. [2002]: *Knowledge and Its Limits*, Oxford: Oxford University Press.

Wixted, J. T. [2020]: 'The Forgotten History of Signal Detection Theory', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **46**, pp. 201–33.

Wu, W. [2014]: *Attention*, New York: Routledge.

Zylberberg, A., Barttfeld, P. and Sigman, M. [2012]: 'The Construction of Confidence in a Perceptual Decision', *Frontiers in Integrative Neuroscience*, **6**, available at <doi.org/10.3389/fnint.2012.00079>.