Forthcoming in *The British Journal for the Philosophy of Science*

Introspection Is Signal Detection

Jorge Morales
Johns Hopkins University

ABSTRACT

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 judgments 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 judgments 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 modeling 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—from signal detection theory (SDT) in particular—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 framework can handle a wider range of cases (both successes and failures) by appealing to a single machinery whose fundamental underlying operation is shared by other cognitive capacities (e.g. perception, decision-making, etc.). In addition to calibrating introspection's whole range of reliability, iSDT also offers a unifying way of understanding response bias and confidence in introspective judgments. 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 & Loettgers, 2016) or, more generally, *model migration* (Lin, 2018). 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 […]. *[T]hey offer resources for further*

investigation and new theoretical insights" (Knuuttila & Loettgers, 2016, pp. 298; my emphasis).

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 & Price, 1973) or the extension of tools developed for understanding the random motion of suspended particles for modeling 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 (Macmillan & Creelman, 2005; Tanner, 1954). SDT has since 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 modeling both in philosophy and psychology.

SDT models perception as the joint outcome of a perceptual discrimination and decision-making. An observer's ability to discriminate stimuli (i.e. their perceptual sensitivity) is proportional to the strength of the perceptual evidence (also known as the internal perceptual response, or perceptual response for short), which in turn tends to be proportional to the strength of the stimuli they encounter. Furthermore, observers need to set a criterion or response rule (also known as response bias) to determine the level of internal response required for classifying a stimulus one way or another. Confidence in one's perceptual judgments can be modeled as further criteria that further classifies internal responses in, say, low and high confidence perceptual judgments. Thus, everything else being equal, an observer is more likely to perceive accurately a person in an alley when the alley is well-lit (i.e. when the stimulus is intense and thereby generates a strong perceptual response) than when the person is in a dark alley (i.e. when the stimulus is weak and generates a weak perceptual response). Moreover, whether the same amount of perceptual evidence leads to perceiving or not someone in the alley depends on how liberal or conservative the criterion is. Similarly, depending on one's confidence criteria, the exact

same perceptual evidence could lead to a perceptual judgement with high or low confidence.[1]

The view I introduce here—iSDT—models introspection similarly to how SDT models perception. Accordingly, iSDT models introspective judgments 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. In a nutshell, iSDT proposes that, everything else being equal, an introspector is more likely to introspect accurately an intense experience (e.g. a strong pain, a vivid mental image, etc.) than a weak experience (e.g. a weak pain, a faint mental image, etc.). iSDT also relies on introspective response bias to fully account for introspective judgments. For example, for an identical weak experience, a liberal introspector may judge they are undergoing that experience (e.g. a weak pain, a faint mental image, etc.) while a conservative introspector may not. Introspective judgments may also be made with low or high confidence depending on where confidence criteria are placed and on the strength of the introspective response.

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, including other inner-sense theories. In section 3 I offer an overview of SDT. Section 4 introduces the notion of "mental strength" (i.e. 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 use 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 similar, systematic explanations for different types of empirical results from the scientific study of consciousness.

[1] There is no limit to how many confidence criteria there are. For simplicity, I only consider the binary case in which decisions are made either with low or high confidence.

2. A Garden-Variety Capacity

The problem of calibration "arises for any scientific instrument and cognitive capacity" (Goldman, 2004, p. 14), and introspection is no exception. The need for calibrating introspection suggests two desiderata. First, an adequate theory of introspection should have the right scope, that is, it should explain introspection's full range of reliability. This means that the conditions that favor both accurate and inaccurate introspective judgments 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, it should also 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 theory.

Prima facie, a reasonable assumption to begin 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.

---

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

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

As I will understand it here, introspection is the focusing of one's attention on one's current conscious experiences to make judgments 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 (e.g. directing their attention in the right time and manner). Thus, introspective judgments 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 (e.g. experiencing a whole visual field, including both central and peripheral regions) that we do not (fully) introspect (e.g. one does not always direct attention towards, and makes introspective judgments about, peripheral vision).[5] In any case, iSDT will try to capture *these* type of introspective judgments. Relatedly, 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 (e.g. pain, mental images and perceptual experiences).[6]

2.1 Infallibility & unreliability

The GARDEN-VARIETY assumption and this way of understanding introspective judgments are in clear tension with prominent views that take introspection to be either infallible (self-intimating, transparent, or privileged and impervious to error in some other way). This is true too for views that consider introspection to be completely unreliable. Both kinds of views bypass the problem of calibration: if

[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: (Carruthers, 2000; Chalmers, 2010; Giustina & Kriegel, 2017; Goldman, 2006; Hatfield, 2005; Peacocke, 1998; Rosenthal, 2005; Ryle, 2009; Schwitzgebel, 2012; Wu, 2014).

[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 the intensity of conscious experiences is linked to the accuracy of introspective judgments.

[6] In principle, iSDT could be extended to other conscious experiences as long as they have an intensity dimension (e.g. emotions, moods, and perhaps occurrent thoughts and desires).

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, 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" (Descartes, 1984, AT VII 29). More recently, Gertler argues that introspection takes place via pure demonstrative reference achieved via 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>." (Gertler, 2001, p. 321) Phenomenal contents are supposed to be embedded in the introspective judgment "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 GARDEN-VARIETY, 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, pp. 323; my emphasis) Several other defenses of some sort of introspective infallibility—especially about occurrent phenomenal experiences—abound in the recent literature (Chalmers, 2003; Horgan & Kriegel, 2007; e.g. Shoemaker, 1996).

My goal here is not to discuss at length these views.[7] 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 judgments 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 even "I'm experiencing this". As Schwitzgebel correctly points out: "there is a reason optimists like the example of pain and foveal visual experience of a single bright color. It is hard, seemingly, to go too badly wrong in introspecting really vivid, canonical pains and foveal colors. But to use these cases only as one's inference base rigs the game." (Schwitzgebel, 2008, pp. 259-60) Once more complex (yet completely common) cases are considered, the infallibility of introspection becomes much harder to maintain.

[7] See section 2.3 for a brief discussion of Shoemaker in the context of his criticism of inner-sense theories.

This acute observation about this "diet" of examples, however, need not turn us into skeptics about introspection. For instance, Schwitzgebel thinks that "we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or 'phenomenology')" (Schwitzgebel, 2008, p. 247). 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 Introspection as an inner-sense

Departing from the tradition that considers introspection infallible allows iSDT to 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".[8] Theories of inner-sense have many reputable defenders (Armstrong, 1968; Goldman, 2006; Kant, 1998; e.g. Locke, 1975; Lycan, 1996), but this kind of theories has acquired a bad reputation. So bad that philosophers often "find it unpersuasive, even repugnant" (Goldman, 2006, p. 225). Against this trend, iSDT aims to become an attractive option for modeling introspection.

Part of the distaste for inner-sense mechanisms stems from common simplifications by critics and, sometimes, champions too. Armstrong, a notable proponent of inner-sense, compares introspection to *bodily* perception because it happens without a "proper organ" and its object "is private to each perceiver" (Armstrong, 1968, p. 325). While it is true that introspection does not have a proper organ, the comparison is unfortunate. Critics of inner sense sometimes also base their

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

objections on misguided analogies. Hill, 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" (1988, pp. 12-3). Neither of these, however, are adequate points of similarity between perception and introspection. Rather than a literal organ (or lack thereof), it is the *type of internal processing* what makes perception and introspection similar.

Hill does raise a criticism against inner-sense views worth considering. 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" (1988, p. 13). Initially, this may indeed appear as a reasonable criticism. However, the inner-sense theorist need not deny that the inner-sense mechanism alters its target states and she does not need to accept either that perceptual processing does not alter its target.[9]

On one hand, it is not generally true that detection mechanisms do not alter their target objects. Measuring an objects' temperature without altering it—even if just slightly—is practically impossible. So, when detecting our experiences we can alter them. For instance, introspecting may make experiences stronger: a pain may become stronger, a mental image may become more vivid, a visual experience may become more intense. This implies that we hardly, if ever, introspect "pure" experiences. Which is, of course, the *right* result (one that Hill himself embraces): we cannot know *exactly* what an unintrospected (e.g. unattended) conscious experience is like (how could we if we are not introspecting them!). If we wanted to say something about "pure" experiences, we would need to rely on memory. It should be obvious that this opens the possibility of 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, once again, between the eye and some internal organ. Rather, it is between *internal* perceptual and introspective *processes*.

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

Introspection can be successfully modeled after perception, but introspection is not perception any more than perception is receiving radio signals, the original domain of application for SDT models. Moreover, orienting our eyes (e.g. foveating) and, more importantly, orienting our attention most definitely alters perceptual *representations* and perceptual *experiences* of an object (Carrasco, 2011; Carrasco, Ling, & Read, 2004). So, the inner-sense theorist can admit that introspective attention affects the target experience, but the critic must admit that a fair comparison with perception would highlight that perceptual attention alters the perceptual processing of the target stimulus.

Shoemaker's (1996) 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." (Shoemaker, 1996, pp. 275; my emphasis) SELF-BLINDNESS, according to Shoemaker, is "nonsense". Instead, he defends SELF-TRANSPARENCY, which holds that, 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.[10]

Inner-sense approaches to introspection deny SELF-TRANSPARENCY and embrace the possibility of SELF-BLINDNESS. In particular, these views allow that target conscious states exist independently from the subject becoming (accurately) aware of them introspectively (e.g. 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. (Hill (1988), Williamson (2002) and Srinivasan (2015), among many others, have offered convincing arguments against claims similar to SELF-TRANSPARENCY.) Rather, here I aim to contextualize some of iSDT's assumptions. Not only *some* kind of SELF-BLINDNESS is empirically possible, we should also take the intensity of the introspected experiences into account in our explanations of introspection.

---

[10] Here I follow Stoljar's (2019) reconstruction of Shoemaker's SELF-TRANSPARENCY claim.

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 severe limitations self-evaluating their performance in those tasks. These effects have been observed in both neuropsychological populations (Fleming, Ryu, Golfinos, & Blackmon, 2014), via causal interventions in neurotypical subjects (Cortese, Amano, Koizumi, Kawato, & Lau, 2016; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010) and via psychophysical manipulations (Koizumi, Maniscalco, & Lau, 2015; Samaha, Barrett, Sheldon, LaRocque, & Postle, 2016; Zylberberg, Barttfeld, & Sigman, 2012). While these effects do not show complete self-blindness, they are not subtle either. For example, Fleming and colleagues (2014) showed a 50% reduction in metacognitive efficiency of confidence ratings about performance in patients with prefrontal cortex lesions. Importantly, these patients saw the stimuli without any troubles 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.

It is important to note that the metacognitive ability to assess one's performance in a perceptual task is technically not identical to introspecting the experiences involved in said task. Notwithstanding the differences, they do not seem to be relevant in practice. Asking subjects to rate confidence in their performance produces virtually identical results to asking subjects to introspect how *visible* the stimulus was (Peters & Lau, 2015). Even though a confidence report may not be a theoretically perfect substitute for an introspective report of 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 task where subjects, even though in principle doing a task about an external stimulus, they have to introspect and compare their experiences in order to perform well in the task (Chirimuuta, 2014; Spener, 2015). Thus, although distinct in principle, metacognitive failure in fact provides a window into introspective failure.

Beyond potential malfunctions of the introspective apparatus, reliability under *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, they cannot be appropriately evaluated in the iSDT framework (or any framework that accepts GARDEN-VARIETY). 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 present the building blocks of iSDT's framework for thinking about introspection. The framework has a wide scope (i.e. it explains success and failure) and it is explanatorily illuminating (i.e. 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. SDT

Consider the next scenarios, which assume a man is in an alley and 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 the dumpster. His face looks bright and the contours of his facial features 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 ill-defined. It is hard for you to see him. You, mistakenly, categorize his face as just a shadow and judge that the alley is empty. However, you are not confident.

> **DARK ALLEY + NEWS**: Identical to DARK ALLEY except that you heard that a robber is on the run in the neighborhood. 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 DARK ALLEY. The only difference here is that knowing about the robber changes how you categorize the same evidence, thereby changing your perceptual judgment. You are still not confident about what you see.

These scenarios illustrate three paradigmatic features of perception that SDT successfully models: sensitivity, response bias & confidence. In a nutshell, according to SDT, perceptual judgments are determined by sensory sensitivity (i.e. the ability to discriminate stimuli based on the way these shape a psychological decision space) and by response biases (i.e. the manner in which the psychological space is partitioned to generate possible responses) (Macmillan & Creelman, 2005). Paradigmatically, perceiving is modeled as an observer deciding whether an internal perceptual response *p* was generated by a stimulus class S1 (e.g. "stimulus absent", "line oriented left", etc.) or S2 (e.g. "stimulus present", "line oriented right", etc.). 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 the 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 (Figure 1).

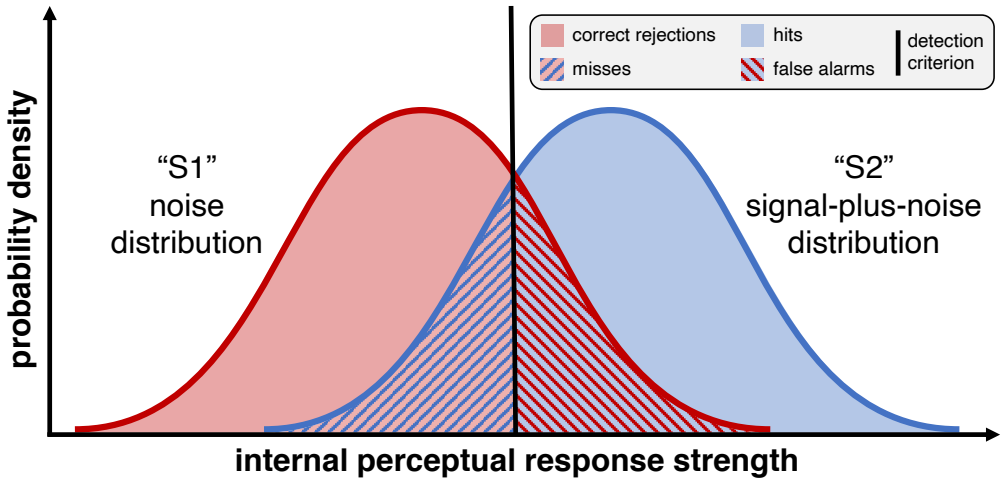

Figure 1. Basic signal detection theory (SDT) model.

3.1 Sensitivity

The average strength of the perceptual response *p* is higher for S2 trials than for S1 stimulus trials. This is reflected in the higher mean of the signal-plus distribution compared to the noise distribution. However, the distributions always have some degree of overlap. The distance between the means of the distribution indicates the observer's sensitivity. The less the distributions overlap, the easier it is for the observer to discriminate S1 from S2. For instance, sensitivity is higher in the BRIGHT ALLEY scenario (Figure 2A) than in the DARK ALLEY and the DARK ALLEY + NEWS scenarios (Figure 2B & 2C). Moreover, for a given trial, the stronger the signal-to-noise ratio of the internal perceptual response *p*, the easier it is for the subject to discriminate signal from noise.

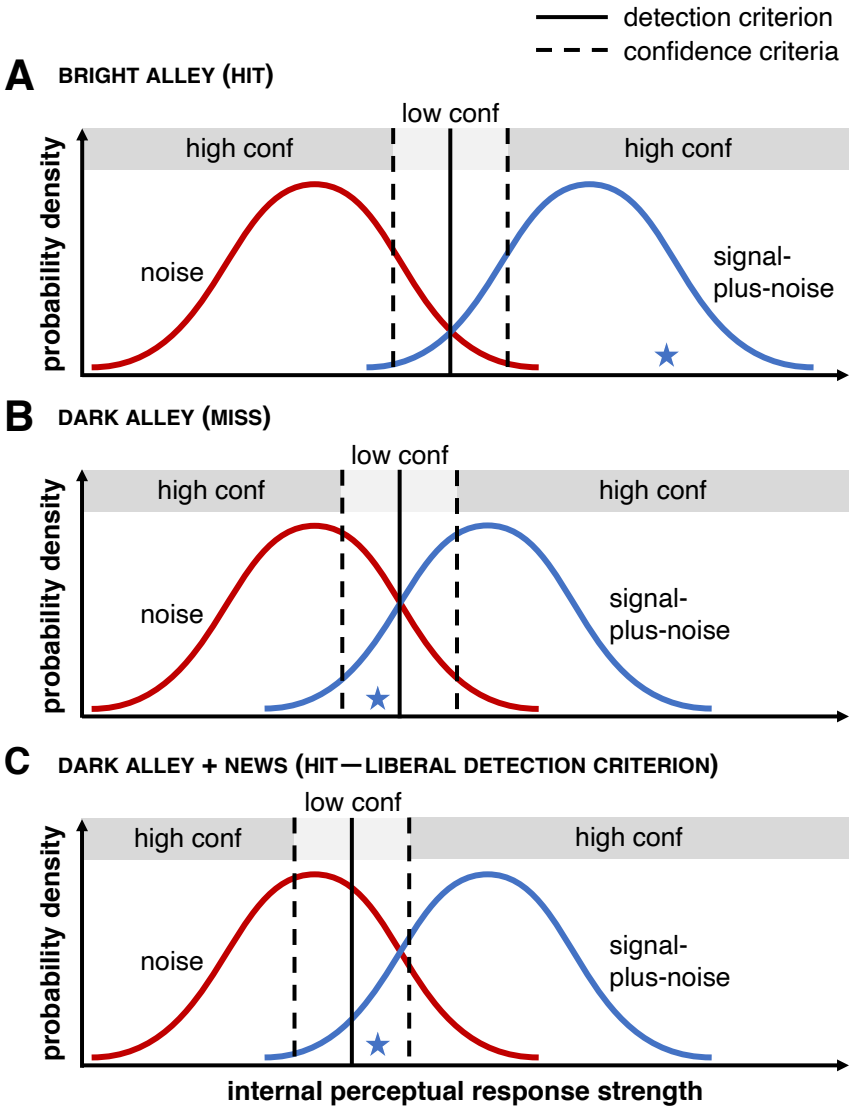

Figure 2. SDT models of the dark alley scenarios.
The distance between the red (noise) and blue (signal-plus-noise) curves represents the observer's perceptual sensitivity (*d'*). **A. BRIGHT ALLEY**. The internal perceptual response (star) produced by the stimulus (i.e. the man) crosses both the detection criterion and the right confidence criterion producing an accurate and confident judgment. **B. DARK ALLEY**. The observer inaccurately judges the alley as being empty, albeit with low confidence because the perceptual response crosses the first confidence criterion but not the detection criterion. **C. DARK ALLEY + NEWS**. An identical perceptual response with an identical sensitivity as in DARK ALLEY produces an accurate detection (still with low confidence) due to more liberal criteria.

A huge advantage of a sensitivity measure such as *d'* is that it is calculated by considering both hits and false alarms (Figure 1), rather than just a raw percentage of undistinguished correct responses. One way of never missing a robber in an alley is classifying every shadow as man. This would ensure every robber is perceived, but one would also be calling 911 unnecessarily all the time. An ideal observer detects the signal every time it is present *and* rejects that the signal is present every time it is absent. A simple raw percentage correct measure mixes both types of correct responses, obscuring the true sensitivity of a subject.[11]

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 judgment, observers classify p as S2 if it exceeds a response criterion c (Figure 1; solid line), and as S1 otherwise. Importantly, whereas sensitivity is a function of stimulus properties and perceptual processing (typically) beyond observer's control, c reflects a response strategy determined by the observer's priors, preferences, goals and other traits (e.g. maximizing the probability of responding correctly, maximizing rewards, degree of risk aversion, perceptual biases, etc.).

Importantly, as the DARK ALLEY and the DARK ALLEY + NEWS scenarios illustrate, sensitivity and response bias are independent from each other (Figure 2B & 2C). While preserving identical sensitivity (i.e. the distance between the distributions' means are the same), an identical perceptual response can yield different perceptual judgments due to changes in the detection criterion. In DARK ALLEY + NEWS, the criterion for detecting the presence of people becomes more liberal (Figure 2C).

---

[11] Observers with identical percentage correct responses can have different *d'*, and hence, different true sensitivities. In an experiment with 100 stimulus-present trials and 100 stimulus-absent trials, Observer 1 correctly detects 70% of present stimuli (i.e. 0.7 hit rate [H]) and correctly rejects 70% of absent trials (i.e. 0.7 correct rejection rate, equivalent to 0.3 false-alarm rate [FA]). Observer 2 correctly detects 90% of present stimuli (0.9 H) and correctly rejects 50% of absent stimuli (0.5 FA). Both observers have 140 (i.e. 70%) correct responses. However, Observer 1's sensitivity is lower than Observer 2's, since hit and false-alarm rates yield d'= 1.05 for Observer 1 and d'=1.28 for Observer 2. The equation for estimating *d'* is:

$$d' = z(H) - z(FA)$$

where z(H) and z(FA) are the z-scores (i.e. the standard deviation units) of H and FA, respectively.

This criterion change results in changes in response accuracy (even if overall sensitivity remains the same): a correct classification in DARK ALLEY + NEWS (hit) and an incorrect classification in DARK ALLEY (miss).

3.3 Confidence

Perceiving, or more specifically, classifying perceptual evidence *p* as S1 or S2, always involves some degree of uncertainty.[12] Confidence in one's perception can also be characterized as resulting from a criterion-setting process (Figure 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 (Figure 2A). If the perceptual response crosses the detection criterion but fails to cross the confidence criterion, observers report detecting the target but with low confidence (Figure 2C). An analogous explanation in the other direction applies too. Observers will report not detecting the stimulus with low confidence when the perceptual response crosses the left confidence criterion, but not the detection criterion (Figure 2B). When the internal response is too weak to cross any criteria, observers will judge with high confidence that the stimulus is absent.

This brief introduction to SDT highlights the crucial role the internal perceptual response plays in modeling perceptual judgments. 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 where introspective responses may stem from: conscious experience intensity or mental strength.

---

[12] Whether this uncertainty is reflected in subjects' phenomenology is a matter of current debate. Recently, the question of whether there is perceptual confidence (i.e. phenomenology of confidence in perceptual experiences), or more generally whether perceptual experiences reflect the probabilistic nature of perception, has been widely discussed (Beck, 2019; Block, 2018; Gross, 2020; Morrison, 2016; 2017; Munton, 2016; Siegel, 2021). Here I am neutral as to whether perceptual phenomenology reflects perceptual confidence or not. Confidence *judgments* may be based on perceptual confidence but the SDT apparatus does not require it.

4. Mental Strength & Introspective Internal Response

The targets of perception (i.e. 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, etc. Stimuli in each modality may be strong (or weak) along more than one dimension (e.g. visual stimuli strength depends on brightness, contrast, saturation, etc.). SDT postulates that after hitting our senses, stimuli produce a 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 above, the probability of making an accurate perceptual judgment largely correlates with the strength of the internal perceptual response (Figure 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 modeling 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 judgments. 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 (Figure 3B). iSDT relies on this obvious fact about our conscious experiences to calibrate introspective reliability.[13]

[13] 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 circles 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.

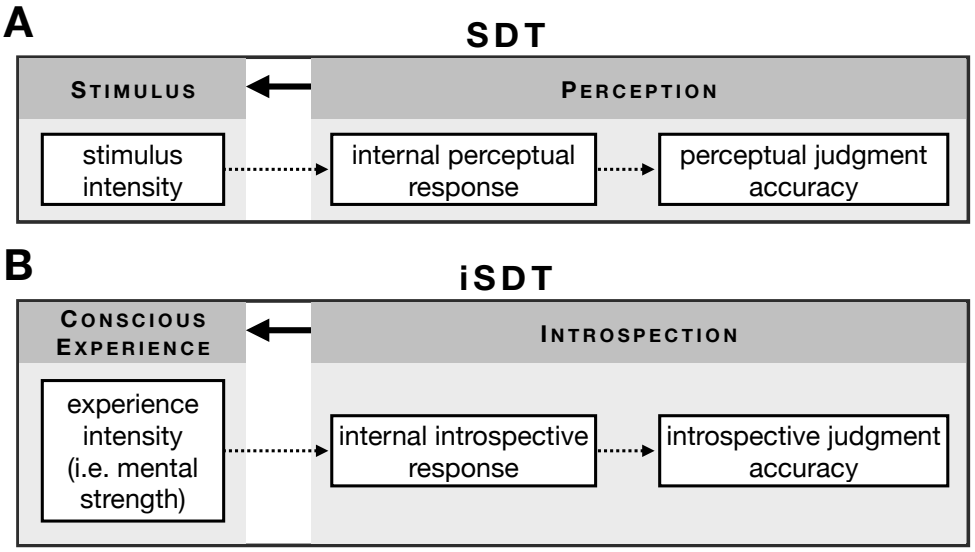

Figure 3. Internal responses in SDT and iSDT.
A. SDT. When perceiving a stimulus, its intensity largely modulates (but does not determine) the strength of the internal perceptual response in the perceiver. Perceptual accuracy is largely driven by the strength of the internal perceptual response (perceptual signal-to-noise ratio). **B. iSDT**. iSDT offers an analogous explanation. When introspecting a conscious experience, its intensity (i.e. its 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.

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.[14] 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 faint one; an experience as of a loud sound typically has more mental strength than an experience as of a quiet sound.

Strong stimuli normally produce experiences with a strong internal response and, in turn, with strong mental strength (and vice versa for weak stimuli) (Peters & Lau, 2015).[15] 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 are, the more intense the visual image tends to be overall.

This correlation, however, does not always hold. Ever-present noise, the subject's overall internal state, deployed attention, familiarity with the stimuli, and many 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 visual image of a very faint candle flame in a dark room lacks clear details by necessity (otherwise it would not be a vivid image with those contents).

The intensity of perceptual experiences can be similarly decoupled from the intensity of the stimuli that generate them. Extreme silence produces intense auditory experiences (Cox, 2014; Sorensen, 2009). A similar decoupling can happen in the visual domain during what is called "subjective inflation" (Knotts, Odegaard, Lau, & Rosenthal, 2019). Peripheral vision is not as sharp as foveal vision, which results in weaker internal responses pertaining peripheral stimuli. Subjects may nonetheless enjoy intense and detailed experiences in the periphery—sometimes even more

---

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

[15] 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 is needed is that there is a rough correlation in normal cases between stronger internal perceptual responses and more intense conscious experiences. No one disagrees about this general correlation, even though different views disagree about the exact conditions that make this correlation possible or the conditions under which it breaks down.

than in foveal regions (Odegaard, Chang, Lau, & Cheung, 2018). Naturally, these 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 colors, and people are normally confident—in fact, overconfident—about their discrimination capacities. Nevertheless, peripheral perception is in fact drastically impoverished (e.g. color discrimination is rather bad). This is a clear case where 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 (which, thereby, does not have any degree of mental strength at all) (Weiskrantz, 1986).[16]

Relatedly, internal responses with identical signal-to-noise ratios may create experiences with different mental strengths. In a recent experimental paradigm called "matched-performance difference-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; Zylberberg et al., 2012). Matched performance is achieved by matching the signal-to-noise ratio of two stimuli that, nevertheless, differ in their overall energy.[17] These stimuli generate matched internal perceptual responses making equally difficult to discriminate the signal even though the stimulus with more energy *looks* more intense (e.g. the contrast looks more marked). Likely, this increase in mental strength is what makes subjects rate their 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, vivid 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

[16] Note that this is true even if blindsight is reinterpreted as qualitatively degraded conscious vision (Phillips, 2020).

[17] For an example of this kind of stimuli, see (Samaha et al., 2016, Figure 1A) which you can find visiting https://tiny.cc/sampleMPDC.

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 next section for examples.)

The details of a theory of mental strength need not be further specified here. 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.

5. iSDT

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 examples as these have been popular in the introspection literature, and then expand the framework to visual imagery and perceptual experiences in section 6. Consider the following scenarios. The first three assume you are in fact experiencing pain; the last one assumes you are not.[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 seems odd though—of course you are confident you are in such a 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 STRONG PAIN).
>
> **MILD PAIN + FAMILIARITY**: This scenario is identical to MILD PAIN (i.e. by stipulation, the intensity of the pain is identical in both scenarios). Here,

[18] For simplicity, in what follows I discuss pain *detection* (i.e. whether someone is or not in pain), but the iSDT machinery can be equally applied to *discrimination* (e.g. whether this pain is throbbing or stabbing).

> however, you are familiar with dental pains. And because you have experienced them before, you know it usually takes several hours for the pain killers to make them go completely away. This time you accurately introspect that you are still in pain but you are not very confident.
>
> **DENTAL FEAR**: You go to the dentist for a routine clean up. You cannot experience any pain because you are under a powerful local anesthetic. 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. At some point, you yelp and report feeling an intense pain. The dentist is confused: they have not even touched you.[19]

STRONG PAIN—and to some extent MILD PAIN + FAMILIARITY—seems quite plausible, but, admittedly, MILD PAIN and DENTAL FEAR may strike some as counterintuitive: how could you be in pain and miss it, or 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, unlike perceptual errors which can easily be pointed out by someone else (Alston, 1971; Dennett, 2002; Langland-Hassan, 2017; Rorty, 1970). Thus, intuition, introspection, or the lack of correction from others are not good routes to discover introspective errors or whether they are possible. A wide-scope, illuminating theory, in contrast, 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 GARDEN-VARIETY—is to show how iSDT can explain all of them in a *systematic* and *principled* way.

---

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

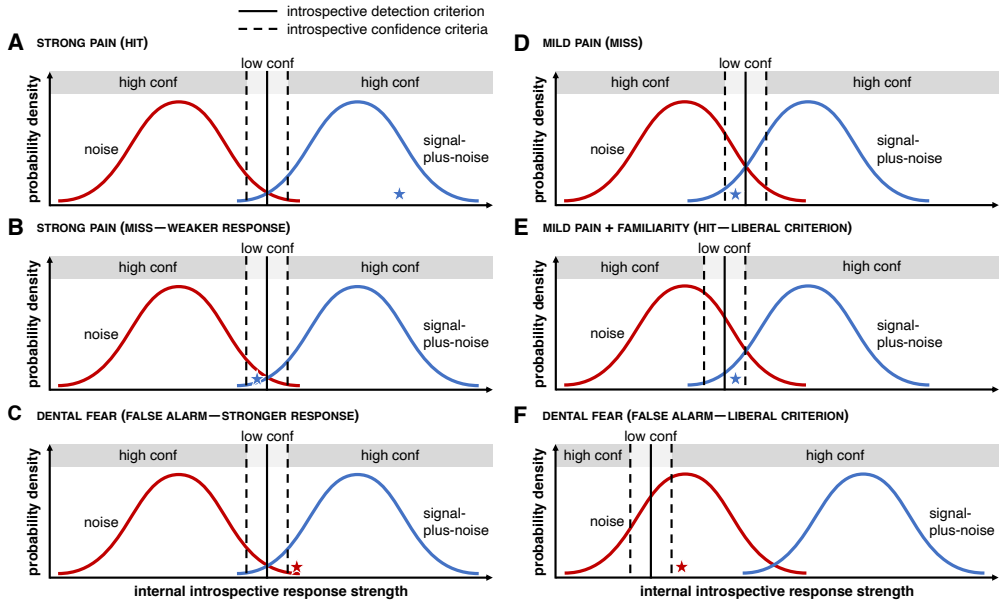

Figure 4. Pain scenarios as modeled by iSDT.

A. STRONG PAIN (HIT). The barely overlapping-distributions indicate high sensitivity; a strong introspective response *i* (blue star indicating *i* is drawn from the signal-to-noise distribution) is accurately and confidently classified as pain. **B. STRONG PAIN (MISS).** Random factors that weaken the introspective response during an identical strong pain result in inaccurate 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 (red 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)**. 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 these kind of pains. Here, a mild pain with introspective evidence *i* (blue star) is missed but with low confidence. **E. MILD PAIN + FAMILIARITY (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 (i.e. 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 (i.e. pain) distribution.

Similarly to how perceptual judgments are conceived in SDT, introspective judgments in iSDT are determined by sensitivity (i.e. the ability to discriminate experiences based on the way these shape a psychological decision space) and by response biases (i.e. the manner in which the psychological space is partitioned to generate possible responses). Introspecting is modeled as an introspector deciding whether an internal introspective response *i* was generated by a conscious-experience class C1 (e.g. "pain absent", "burning pain", etc.) or C2 (e.g. "pain present", "stabbing pain", etc.). The introspective response corresponds to the strength of the introspective evidence, in turn modulated by the intensity of the conscious experience (i.e. 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 (Figure 4).

5.1 Introspective sensitivity

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

Note that the probability of error in STRONG PAIN is very small (i.e. the area of overlap between the two curves in Figure 4A). Introspective errors under these conditions should be quite rare, but not impossible (in the same way 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 (Figure 4B). This could happen even if the mental strength of a particular painful experience is strong. Recall that the relation between mental strength and introspective response is not deterministic (Figure 3B). The process that gives rise to an introspective response from a strong experience may get corrupted, generating a weak introspective response (i.e. weaker

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

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 as of a strong pain are possible too (Figure 4C). The iSDT framework has a straightforward way of accommodating rare cases such as DENTAL FEAR. Patients' fear of the procedure in conjunction with vibrations produced by the dentist's instruments (i.e. noise) may significantly increase the introspective response. This, in turn, produces a pain report even though no pain is (could be) experienced under the circumstances (i.e. under potent anesthesia or without dental nerves). (See next subsection for an explanation of DENTAL FEAR 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 MILD PAIN and MILD PAIN + FAMILIARITY is lower. This can be modeled by lowering the mean of the signal-plus-noise distribution making the distributions overlap more (Figures 4D & 4E). 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.

5.2 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 judgments may arise not (only) because of an insensitive or inaccurate machinery, they can also take place 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 MILD PAIN and MILD PAIN + FAMILIARITY, introspective sensitivity and internal responses are, by stipulation, identical (Figure 4D & 4E).

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

And yet, in the former scenario the pain is not reported while it is in the latter. iSDT model these scenarios by shifting the introspective criterion. In MILD PAIN + FAMILIARITY, knowledge about the time course of the pain killers 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 MILD PAIN you have an unbiased detection criterion, and the same weak introspective response is insufficient for introspecting that you are in pain. Finally, the DENTAL FEAR scenario could also be explained by a liberal shift in the criterion (instead of an increased introspective response as suggested above; Figure 4C). According to this explanation, the fearful patient is more willing to classify as pain a really weak introspective response produced by vibrations—i.e. noise (Figure 4F). 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 (section 3.1). 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 judgments. One consequence of having a liberal criterion in MILD PAIN + FAMILIARITY is not only that existing mild pains are more easily detected (higher hit rate), it also implies 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) 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 behavior in quite a subtle way (perhaps subtler than some current philosophical approaches to introspection allow), making introspection's machinery consistent with that of other faculties.

5.3 Introspective confidence

Infallibilists and skeptics 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 judgment is a function of the strength of the introspective response and the placement of confidence criteria (dashed lines in Figure 4). To capture the common intuition that introspection is, if not infallible, unlikely to be significantly wrong most of the time (pace Schwitzgebel and other skeptics), confidence criteria in Figure 4 are placed much closer to the detection criterion than they were in Figure 2 in the perceptual case. This entails introspective responses stay in the high-confidence regions of the introspective decision axis in most cases (e.g. in STRONG PAIN [Figure 4A], which I take to be a representative case of many of our introspective judgments).

Importantly, at least sometimes introspective judgments are made with low confidence, a fact that skeptics have used to mount their generalized attack on introspective reliability (e.g. MILD PAIN and MILD PAIN + FAMILIARITY [Figures 4D & 4E]). Low confidence, however, is not a perfect guide to learning the degree of 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 judgments 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 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 behavior such as response biases and confidence. Finally, an important lesson modeling these cases can offer is how important

[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).

it is to let theory—not intuition or introspection—lead the way we reason through complex and, at least *prima facie*, unintuitive scenarios.

6. Beyond Pains: Mental Imagery and Perception

iSDT models introspection in the same way for any conscious experience with a degree of intensity (i.e. mental strength) that generates an introspective response.[23] Here I can only briefly sketch how to expand iSDT to mental imagery and visual perception. Rather than a full treatment of iSDT application 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 in multiple kinds of conscious experience.

A particularly vivid mental image of a simple object (e.g. a red apple) is an example of an intense experience in the imagery domain. iSDT predicts cases like this to produce strong introspective responses that result in accurate, confident introspective judgments. 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 judgments with lower confidence can be expected. It is a matter of contention what makes an image more or less vivid, or even what is meant by vividness (Kind, 2017). However, it is reasonable to assume that overall mental strength of a mental image is the result of an aggregate of intensity across several dimensions; for example: sensory properties (e.g. brightness, loudness, etc.), clarity, number and salience of details, the feeling of presence of the imagined objects or events, and the overall stability of the image (Cornoldi, De Beni, Giusberti, & Marucci, 1991).

Perceptual experiences can be expected to follow exactly the same mold. As discussed above, there is a strong link between perceptual response and mental strength and between mental strength and introspective response, which predicts we are likely to introspect accurately strong experiences of strong stimuli. But this

[23] Insofar as other experiences have degrees of intensity they can also be modeled by iSDT (e.g. itches, emotions, moods, action-awareness, sense of bodily ownership, and perhaps even thoughts and desires).

link 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) and Sperling-like studies (Landman, Spekreijse, & Lamme, 2003; Sligte, Scholte, & Lamme, 2008; Sperling, 1960) 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 would not reflect the nature of external stimuli or 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, subjects do not fail to access their phenomenal contents and they do not introspect their experiences inaccurately. Rather, subjects accurately introspect rich (or enriched) experiences that are, nevertheless, dissociated from perceptual processing (see Knotts et al., 2019 for a similar approach). It is poor perceptual processing what 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 inaccurate introspection. 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, which have been successfully modeled 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 modeling variations in introspective sensitivity, response bias, and confidence.

References

Alston, W. (1971). Varieties of Privileged Access. *American Philosophical Quarterly*, *8*(3), 223–241.

Armstrong, D. M. (1968). A Materialist Theory of the Mind. London: Routledge & Kegan Paul.

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

Beck, J. (2019). On Perceptual Confidence and "Completely Trusting Your Experience." *Analytic Philosophy*, *59*(236), 385–15. http://doi.org/10.1111/phib.12151

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, *30*(5-6), 481–548. http://doi.org/10.1017/S0140525X07002786

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170341–10. http://doi.org/10.1098/rstb.2017.0341

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. http://doi.org/10.1016/j.visres.2011.04.012

Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, *7*(3), 308–313. http://doi.org/10.1038/nn1194

Carruthers, P. (2000). Phenomenal Consciousness. New York: Cambridge University Press.

Chalmers, D. J. (2003). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives* (pp. 220–272). New York: Oxford University Press.

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*(5), 914–926. http://doi.org/10.1086/677890

Cornoldi, C., De Beni, R., Giusberti, F., & Marucci, E. (1991). The study of vividness of images. In R. H. Logie (Ed.), *Mental Images in Human Cognition*. Elsevier.

Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, *7*, 13669. http://doi.org/10.1038/ncomms13669

Cox, T. (2014). The Sound Book. New York: W. W. Norton & Company.

Dennett, D. C. (2002). How could I be wrong? How wrong could I be? *Journal of Consciousness Studies*, *5-6*, 13–16.

Descartes, R. (1984). The Philosophical Writings of Descartes. (J. Cottingham, R. Stoothoff, & D. Murdoch, Eds.) (Vol. I & II). Cambridge: Cambridge University Press.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*(1), 3–25. http://doi.org/10.3758/bf03196254

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-Specific Impairment in Metacognitive Accuracy Following Anterior Prefrontal Lesions. *Brain*, *137*(10), 2811–2822. http://doi.org/10.1093/brain/awu221

Gertler, B. (2001). Introspecting Phenomenal States. *Philosophy and Phenomenological Research*, *63*(2), 305–328.

Giustina, A., & Kriegel, U. (2017). Fact-Introspection, Thing-Introspection, and Inner Awareness. *Review of Philosophy and Psychology*, *8*(11), 143–164. http://doi.org/10.1007/s13164-016-0304-5

Goldman, A. (2004). Epistemology and the Evidential Status of Introspective Reports. *Journal of Consciousness Studies*, *11*(7-8), 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 & Language*, *35*(3), 377–389. http://doi.org/10.1111/mila.12280

Hatfield, G. (2005). Introspective evidence in psychology. In P. Achinstein (Ed.), *Scientific Evidence: Philosophical Theories & Applications* (pp. 259–286). Baltimore: Johns Hopkins University Press.

Hill, C. S. (1988). Introspective awareness of sensations. *Topoi*, *7*(1), 11–24. http://doi.org/10.1007/BF00776205

Hohwy, J. (2011). Phenomenal Variability and Introspective Reliability. *Mind & Language*, *26*(3), 261–286. http://doi.org/10.1111/j.1468-0017.2011.01418.x

Horgan, T., & Kriegel, U. (2007). Phenomenal Epistemology: What Is Consciousness That We May Know It So Well? *Philosophical Issues*, *17*(1), 123–144. http://doi.org/10.1111/j.1533-6077.2007.00126.x

Humphreys, P. (2002). Computational Models. *Philosophy of Science*, *69*(S3), S1–S11.

Kant, I. (1998). Critique of Pure Reason. (P. Guyer & A. W. Wood, Eds.). Cambridge: Cambridge University Press.

Kind, A. (2017). Imaginative Vividness. *Journal of the American Philosophical Association*, *3*(1), 32–50. http://doi.org/10.1017/apa.2017.10

Knotts, J. D., Odegaard, B., Lau, H., & Rosenthal, D. (2019). Subjective inflation: phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, *29*, 49–55. http://doi.org/10.1016/j.copsyc.2018.11.006

Knuuttila, T., & Loettgers, A. (2016). Model templates within and between disciplines: from magnets to gases – and socio-economic systems. *European Journal for Philosophy of Science*, *6*(3), 377–400. http://doi.org/10.1007/s13194-016-0145-1

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295–1306. http://doi.org/10.3758/s13414-015-0843-3

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164.

Langland-Hassan, P. (2017). Pain and Incorrigibility. In J. Corns (Ed.), *The Routledge Handbook of Philosophy of Pain*. London: Routledge.

Lin, C.-H. (2018). Tool Migration: A Framework for Analyzing Cross-disciplinary Use of Mathematical Constructs. *PhilSci Archive*, 1–11. Retrieved from http://philsci-archive.pitt.edu/14929/

Locke, J. (1975). An Essay Concerning Human Understanding. (P. H. Nidditch, Ed.). Oxford: Clarendon Press.

Lycan, W. G. (1996). Consciousness and Experience. Cambridge, MA: MIT Press.

Macmillan, N. A., & Creelman, C. D. (2005). Detection Theory (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Maxwell, J. C. (1861). XXV. 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*(139), 161–175.

Meier, M. L., de Matos, N. M. P., Brügger, M., Ettlin, D. A., Lukic, N., Cheetham, M., et al. (2014). Equal pain—Unequal fear response: enhanced susceptibility of tooth pain to fear conditioning. *Frontiers in Human Neuroscience*, *8*(526), 1–11. http://doi.org/10.1016/0022-510X(94)90239-9

Merton, R. C. (1969). Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case. *The Review of Economics and Statistics*, *51*(3), 247–257. http://doi.org/10.2307/1926560

Morrison, J. (2016). Perceptual Confidence. *Analytic Philosophy*, *57*(1), 15–48.

Morrison, J. (2017). Perceptual Confidence and Categorization. *Analytic Philosophy*, *58*(1), 71–85.

Munton, J. (2016). Visual Confidences and Direct Perceptual Justification. *Philosophical Topics*, *44*(2), 301–326. http://doi.org/10.5840/philtopics201644225

Odegaard, B., Chang, M. Y., Lau, H., & 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: Biological Sciences*, *373*(1755), 20170345–10. http://doi.org/10.1098/rstb.2017.0345

Peacocke, C. (1998). Nonconceptual Content Defended. *Philosophy and Phenomenological Research*, *58*(2), 381–388.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*. http://doi.org/10.7554/eLife.09651

Phillips, I. (2020). Blindsight is qualitatively degraded conscious vision. *Psychological Review*, 1–100. http://doi.org/10.1037/rev0000254

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*(e223), 1–66. http://doi.org/10.1017/S0140525X18000936

Renero, A. (2019). Modes of Introspective Access: a Pluralist Approach. *Philosophia*, *47*, 823–844. http://doi.org/10.1007/s11406-018-9989-2

Reuter, K. (2011). Distinguishing the Appearance from the Reality of Pain. *Journal of Consciousness Studies*.

Rorty, R. (1970). Incorrigibility as the Mark of the Mental. *The Journal of Philosophy*, *67*(12), 399–424.

Rosenthal, D. (2005). Consciousness and Mind. New York: Oxford University Press.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175. http://doi.org/10.1080/17588921003632529

Ryle, G. (2009). The Concept of Mind. New York: Routledge.
Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating Perceptual Confidence from Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory. *Frontiers in Psychology*, *7*(938), 166. http://doi.org/10.3389/fnint.2012.00079
Samuelson, P. A. (1969). Lifetime Portfolio Selection By Dynamic Stochastic Programming. *The Review of Economics and Statistics*, *51*(3), 239–246. http://doi.org/10.2307/1926559
Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review*, *117*(2), 245–273.
Schwitzgebel, E. (2012). Introspection, What? In D. Smithies & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 29–48). New York: Oxford University Press.
Shoemaker, S. S. (1996). The First-Person Perspective and Other Essays. New York: Cambridge University Press.
Siegel, S. (2021). How can perceptual experiences explain uncertainty? *Mind & Language*.
Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are There Multiple Visual Short-Term Memory Stores? *PLoS ONE*, *3*(2), e1699. http://doi.org/10.1371/journal.pone.0001699.g008
Smith, J. M., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18. http://doi.org/10.1038/246015a0
Sorensen, R. (2009). Hearing Silence. In M. Nudds & C. O'Callaghan (Eds.), *Sounds and Perception* (pp. 126–145).
Spener, M. (2015). Calibrating Introspection. *Philosophical Issues*, *25*(1), 300–321. http://doi.org/10.1111/phis.12062
Sperling, G. (1960). The information available in brief visual representations. *Psychological Monographs: General and Applied*, *74*(11), 1–29.
Srinivasan, A. (2015). Are We Luminous? *Philosophy and Phenomenological Research*, *90*(2), 294–319. http://doi.org/10.1111/phpr.12067
Stoljar, D. (2019). Armstrong's Just-so Story about Consciousness. In P. Anstey & D. Braddon-Mitchell (Eds.), *A Materialist Theory of the Mind: 50 Years On*.
Tanner, W. P., Jr. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.
Weiskrantz, L. (1986). Blindsight: A Case Study and Implications. Oxford: Oxford University Press.
Williamson, T. (2002). Knowledge and Its Limits (pp. 1–353). Oxford University Press.
Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233. http://doi.org/10.1037/xlm0000732
Wu, W. (2014). Attention (pp. 1–327). New York: Routledge.
Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*(September), 1–10. http://doi.org/10.3389/fnint.2012.00079