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 of modeling introspection. By using 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). This framework provides the opportunity to calibrate introspection's reliability and to model its cognitive architecture, at the same time that it has the potential of illuminating the underlying computational processes that support introspection.

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 elemental trustworthiness of one of our most important capacities? Rather than praising or vilifying introspection, however, we should aim to explicate and to calibrate it (Goldman, 2004; Spener, 2015). In other words, we should provide a theory that can explain *when* and *why* it fails and *when* and *why* it succeeds. Here, I present a new framework for introspection that takes lessons learned from the science of perception and extrapolates them to model the whole range of introspective access we have to our conscious states.

1. From Perception to Introspection

1.1 Model Migration

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). A famous model migration example 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 example of this kind of extension is, without a doubt, signal detection theory (SDT). Originally developed during the first half of the twentieth century as a mathematical framework for evaluating radar performance, SDT was later adapted to explain perceptual sensitivity (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).

According to Knuutila and Loettgers, "the notion of a model template aims to capture the intertwinement of a mathematical structure and associated computational tools with theoretical concepts that, taken together, depict a general mechanism that is potentially applicable to any subject or field displaying particular patterns of interaction" (Knuuttila & Loettgers, 2016, p. 295). Importantly, over and above the 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 […]. As such *they offer resources for further investigation and new theoretical insights*" (Knuuttila & Loettgers, 2016, pp. 298; my emphasis).

Here, I argue that we can take the resources and theoretical insights garnered by SDT and extend them to further our understanding of introspection—a domain that has historically resisted satisfactory modeling both in philosophy and psychology. Rather than a strict extension of the mathematical and computational tools developed within SDT, I will leverage its conceptual and theoretical *insights* to explicate and to calibrate our capacity for introspection. Thus, I will build a framework I call "introspective signal detection theory" (iSDT). The goal of introducing iSDT is to allow us to conceptualize introspection more carefully and in more scientific terms than previous approaches have typically allowed. In brief, by offering a functional analysis of the psychological principles under which introspection operates, a fruitful *sketch* of how to think about its reliability and about the cognitive and neural mechanisms that support it should begin to emerge (Piccinini & Craver, 2011).

1.2 A first approximation to SDT & iSDT

In this subsection, I briefly introduce the basic tenets of SDT and iSDT. I will further explain these in sections 3 and 5, respectively.

PERCEPTION & SDT. According to SDT, to perceive is to discriminate and to decide. Perceptual judgments are made by deciding how to classify a stimulus based on the perceptual internal evidence it produces in an observer. An observer's ability to discriminate stimuli (i.e. the observer's perceptual *sensitivity*) normally is proportional to the strength of their internal perceptual responses, which in turn are proportional to the strength of the stimuli they are presented with. Or to be slightly more precise, perceptual sensitivity is proportional to the signal-to-noise ratio of internal perceptual responses. For instance, all things being equal, an observer is more likely to accurately perceive a person in an alley when the alley is well-lit (i.e. when the stimulus is strong) than when the person is in a dark alley (i.e. when the stimulus is weak). This is so because the signal-to-noise ratio of the internal perceptual responses produced in the observer over time are stronger when the person is illuminated than when they are not.

The strength of the internal perceptual response, however, is in itself insufficient for producing a perceptual judgement. An observer needs, in addition to the internal response, a criterion or response rule (also known as a response bias) that determines the level of internal response that is required for making a perceptual judgment one way or another. For instance, an arbitrary internal evidence *e* when looking at someone in a well-lit alley could be classified as "there's a person in the alley" under a neutral or liberal criterion, but the exact same amount of internal evidence *e* could be classified as "there's no one in the alley" under a conservative criterion. Together, the strength of the internal response and the response bias for classifying it are the foundation of this powerful, yet simple, framework for modeling perception. Moreover, it is worth noting that SDT is also effective for modeling other features related to perceptual judgments. In particular, SDT is apt for modeling confidence in one's perceptual decision.

INTROSPECTION. At a general level of description, introspection is the process of focusing one's attention on one's current conscious mental states or mental events to make judgments about them.[1] Accordingly, we can introspect perceptual experiences, pains, mental images, emotions, desires, and thoughts, among others. Introspection thus understood implies *some* amount of effort from the introspector. While it need not be special event, the way I will understand introspection makes it some (minimal) kind of cognitive achievement. At the same time, this means that introspection is independent from the conscious states it targets. For instance, the way I understand introspection implies that at least sometimes we undergo

[1] Many philosophers who, despite having very different views about introspection in particular and the mind more generally, agree that introspection involves some kind of attention oriented towards conscious mental states or events (Carruthers, 2000; Chalmers, 2010; Giustina & Kriegel, 2017; Goldman, 2006; Hatfield, 2005; Peacocke, 1998; Rosenthal, 2005; Ryle, 2009; Schwitzgebel, 2012; Wu, 2014).

conscious experiences (e.g. experiencing a whole visual field, including not just central vision but also the periphery) that we do not introspect (e.g. one does not always direct attention towards, and makes judgments about, the visual periphery). While I take it to be a reasonable assumption, I admit not everyone agrees with this starting point. I'll come back to this issue in section 2.4 when discussing Shoemaker's criticism of inner-sense theories. Relatedly, a theory of 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, the framework I will be putting forward focuses only on conscious sensory experiences (e.g. pain, mental images and perceptual experiences). I expect, but I will not have the space to show, that this framework can be extended to other experiences such as emotions or moods and to occurrent thoughts and desires.

iSDT. An evident fact about conscious experiences is that they have degrees of intensity: pains can be stronger or weaker, mental images can be more or less vivid, perceptions can be more or less striking. I refer to the intensity of conscious experiences as their *mental strength*, and it will play a crucial role calibrating introspective reliability (I provide more details about mental strength in section 4).

The central tenet of iSDT is that the intensity of our conscious experiences (i.e. their mental strength) modulates introspective sensitivity in a similar fashion to the way stimulus strength modulates perceptual sensitivity.[2] For example, everything else being equal, an intense conscious experience is more likely to generate a stronger internal introspective response than a weak experience and, in consequence, it is more likely to be introspected accurately than a weak experience. Concretely, this means introspection of a strong pain, a vivid mental image, or a striking sound is more likely to be accurate than introspection of a mild pain, a weak mental image or a soft sound. iSDT can also model introspective response bias as well as introspective confidence (I develop iSDT in section 5 and discuss some differences between perception and introspection in section 6).

Before developing the view, in the next section I discuss a starting assumption at the foundation of the iSDT framework and contrast it to some alternative views about introspection.

[2] This means that iSDT is presented at the level of the phenomenology of our conscious experiences and, in consequence, it should be understood as a framework that explains introspection at the level of conscious agents. To reiterate a point I made earlier, the parallel with SDT here is taken to provide *insights* from perception rather than a precise mathematical framework for understanding introspection. That said, this approach at the phenomenal level still favors a naturalistic understanding of introspection as a mechanism for the detection of (introspective) signals that has the potential of illuminating introspection's underlying computational and neural processes.

2. Introspection: A Garden Variety Capacity

2.1 Assumptions & desiderata

To provide a naturalistic, mechanistic-based explanation of introspection, I start off with the *prima facie* reasonable assumption that introspection is not unlike the rest of our cognitive capacities. In particular, I assume introspection is not equally reliable across all conditions. Call this the assumption that introspection is a GARDEN-VARIETY capacity. Like all our other faculties, introspection may get things right sometimes and it may get them wrong some other times. On one hand, introspecting a simple, clear experience, of a simple feature, under ideal conditions, is likely to yield accurate judgments most of the time. On the other hand, it is plausible that introspection—insofar as it is a human capacity—gets things wrong sometimes, especially when conditions are suboptimal, we are distracted, or the target experience is faint or complex in some way.

I take these facts to be important *explananda* for any theory of introspection. A theory of introspection should, then, be able to explain the conditions in which it is reliable (Goldman, 2004; Spener, 2015). In other words, "a crucial problem for the theory of introspection is to fix its range of reliability" (Goldman 2004, p. 14). As Goldman correctly points out, this problem of *calibration* "arises for any scientific instrument and cognitive capacity" (*idem*; my emphasis). The contents and conditions of operation of introspection are clear: as the definition provided in the previous section indicates, introspection is limited to attended conscious mental states and events.[3] However, to calibrate introspection we need to specify exactly what makes it fail and succeed.

The problem of calibration suggests two *desiderata* for a theory of introspection. First, an adequate theory of introspection should have the right *scope*. This means that the conditions that favor both accurate and inaccurate introspective judgments should be covered by the theory. The theorist's job, then, is to provide a characterization of the full range of the reliability of introspection.

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. Ideally, this explanation should be the same, or of the same kind, for successes and failures. Moreover, everything else being equal, the theory should

[3] Attempts to introspect unconscious cognitive processes inevitably result in illegitimate introspective-like judgments (Nisbett & Wilson, 1977).

appeal to a single mechanism or explanation that covers both success and failures.[4] Consider perception and SDT as a successful example of a theory that explains a capacity's full range of reliability in an illuminating way. SDT explains sensitivity by appealing to the signal-to-noise ratio of the internal perceptual response, which has the power to explain perceptual sensitivity's whole range: from chance performance to ceiling, easy and hard cases are explained (even predicted!) by the same principle. Similarly, a theory of introspection should explain (and predict) when accurate and inaccurate introspection is likely to happen. Thus, accurate and inaccurate introspective judgments are a live possibility based on these theoretical considerations (no matter how intuitive or unintuitive this possibility may seem).

2.2 Infallibility?

Admittedly, introspection has not always been considered a garden variety capacity. In fact, the notion that introspection is infallible, "self-intimating", transparent, or in some other way privileged and impervious to error, has 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, 1985, 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: it is thus [here, now]. "By appropriately attending to the dull throbbing sensation [of a headache], you demonstratively pick out the phenomenal content <dull throbbing>." (Gertler, 2001, p. 321) Thus, according to her, phenomenal content is embedded in the introspective judgment 'it is thus here and now', preventing any sort of error when introspecting one's phenomenally conscious states. Moreover, according to her, the way introspection works is unlike any other *mechanism* the mind uses to process information. Rather, "pure demonstrative reference allows the subject to grasp the content directly […] in the sense that there is no causal gap between the referring state and its referent, the phenomenal content. For the referring state instantiates the phenomenal content, by virtue of embedding its token." (Gertler, 2001, p. 323) Her view is simply at the antipodes of what iSDT assumes and proposes. Several other defenses of some sort of introspective infallibility abound in the recent literature (e.g. Shoemaker 1996; Chalmers 2003; Horgan and Kriegel, 2007).

[4] An implication of this *ceteris paribus* clause is that pluralist accounts aren't immediately precluded (Renero, 2019; e.g. Schwitzgebel, 2012). Rather, in providing the details of iSDT, its wide and illuminating scope makes it unnecessary to appeal to multiple mechanisms or faculties of introspection to accommodate the cases I focus on.

Discussing in detail any of these views, let alone all of them, would take too much space and it is not my main goal in this article.[5] However, it's worth noting that introspective infallibility is often defended based on a very limited set of examples.[6] It might be tempting to think introspective judgments are always accurate if the examples one relies on are always of the type "I'm in pain now" or "I am seeing a red tomato" or even "I'm experiencing *this*".[7] As Schwitzgebel correctly points out: "[T]here's 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." (2008, pp. 259-60) Once more complex yet completely common cases are considered, the infallibility of introspection seems much harder to maintain.[8]

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')" (2008, p. 247). Rather than embracing this other extreme, what we need is a principled method for calibrating the whole range of reliability of our capacity to introspect. The GARDEN VARIETY assumptions should make us find it equally implausible that introspection is infallible and that it is, at least during normal circumstances, utterly broken. Just as we try to understand how perception or memory or decision-making works when they do and how 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.

2.3 Uniqueness?

Related to introspection's privileged nature, there is also a long history defending its *peculiarity* or *uniqueness*. This is not a coincidence. Having an infallible capacity *requires* it to be of a different kind from the rest of our all-too-normal, fallible faculties. However, with the

[5] For more extensive treatments of this issue, see e.g. Alston 1971; Schwitzgebel 2019; Gertler 2020.

[6] Another large part that explains the privileged character of introspection is its proposed peculiar nature, which I discuss below.

[7] Some defend the infallibility of introspection *only* for these restricted cases (e.g. Horgan & Kriegel, 2007). On one hand, this might seem reasonable; on the other hand, this opens up the necessity of postulating multiple explanations for different degrees of introspective reliability: one for infallible cases, another one for fallible ones. Everything else being equal, it's more desirable to postulate a single explanation that covers the whole range of reliability.

[8] Compare introspecting a mental image of a red circle against introspecting a mental image of the façade of your childhood home. Unless you have a uniquely vivid imagination, introspecting these two cases are likely to have different success rates. At the very least, the first case should be easier than the second one.

rejection of infallibility goes the need for a special faculty of introspection. As I argue in the following sections, the fundamental traits of introspection (its range of reliability in particular) can be explained by a garden variety detection mechanism.

This initial assumption doesn't entail that introspection is identical or reducible to another capacity (e.g. perception, attention, working memory, decision-making or a combination thereof). Despite the parallels I'll draw with SDT, under iSDT introspection is *not* a type of perception (see section 6 for further discussion). Rather, both perception and introspection are instances of a general kind of signal detection. iSDT starts off the assumption that the basic mechanisms and computations underlying introspection are not radically different from the principles in operation when we perceive the world, attend to it, remember the past, or make decisions.[9] These otherwise very different capacities rely on shared features such as having a wide range of reliability, operating under conditions of noise and uncertainty, evaluating evidence, discriminating internal signals from noise, etc. Of course, introspection's cognitive architecture and computational profile isn't *exactly* like anything else in the mind and, therefore, it is a distinct (although not special) capacity in its own right (in the same way perception and decision-making are distinct even though they both involve, say, evaluating evidence under conditions of uncertainty).

2.4 Introspection as an inner-sense

By departing from the tradition that considers introspection unique, iSDT embraces a different one: a tradition that considers introspection as a kind of "inner-sense" mechanism.[10] While the views that consider introspection to be the outcome of some sort of inner-sense or scanning "device" have reputable defenders (Armstrong, 1968; Goldman, 2006; Kant, 1998; e.g. Locke, 1975; Lycan, 1996), the whole idea of an "inner-sense" has acquired a bad reputation. In fact, philosophers "find it unpersuasive, even repugnant." (Goldman, 2006, p. 225) I hope to make iSDT not only non-repugnant, but even attractive.

[9] The kind of similarity I have in mind is what sometimes is called *canonical computations*: "There is increasing evidence that the brain relies on a set of canonical neural computations, repeating them across brain regions and modalities to apply similar operations to different problems." (Carandini and Heeger, 2011, p. 51) iSDT is neutral about the details of the neural implementation of introspection, but the idea of canonical computations applies at the computational level too: similar fundamental psychological principles of operation are applied to solve different problems in different contexts giving rise to distinct capacities. Here I focus on signal detection as a widespread, general mechanism shared by perception and introspection.

[10] Many other philosophers also depart from 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).

Part of the distaste for inner-sense mechanisms stems from simplification by critics and the admittedly faulty metaphors some defenders have offered (see Picciuto & Carruthers, 2014 for discussion). Armstrong himself, a notable proponent of inner-sense, compares introspection to bodily perception because it happens without a "proper organ" and its object (our own body) "is private to each perceiver" (Armstrong 1968, p. 325). Critics of inner sense views also take literal analogies. Hill (1988), for instance, discusses what he labels as the "inner eye hypothesis," according to which, 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" (pp. 12-3). These, however, are not the relevant points of similarity between perception and introspection that we should focus on. Rather than the literal organ through which we perceive, it is the *type of processing* that brings about perception and introspection that are similar, not some supposed literal similarity between external and inner senses.[11]

But Hill raises another criticism against inner sense that deserves consideration. 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 are mistaken when they argue that "the internal qualities of sensations do not change when one scans them" (Hill 1988, p. 13). But this is not something the inner-sense theorist needs to commit to. The inner-sense theorist need not defend that the detection mechanism doesn't alter the target state. While it might be literally true that the physical eye doesn't alter physical objects, it is not in general true that detection mechanisms don't alter their target objects. Many types of measurement affect the measured object. For example, measuring temperature without thereby altering the temperature of the measured object or surface—even if just slightly—is practically impossible. That one's conscious experiences are altered when they are introspected should rather be part of an inner sense theory. And, in fact, in virtue of introspecting our experiences we often alter them, for example, making them stronger (e.g. an attended pain may become more painful, an introspected mental image may become more vivid, etc.). That we are don't know what a completely unintrospected experience is like seems to be the *right* result (incidentally, we cannot know either what unobserved objects look like—at least not from vision).

The charge that the eye does not alter the internal qualities of its objects seems to miss the mark too. While perhaps literally true,[12] once again, the right comparison between perception

[11] Not all inner-sense views are faulty in this way. Goldman (2006; ch. 9), for example, is a notable exception in that he provides a detailed account of the cognitive processes, mental properties, and brain mechanisms required for a self-detection introspective mechanism.

[12]

and introspection is not between the eye and some internal organ. Rather, it is between perceptual internal *processes* and introspective *processes*. Orienting our eyes (e.g. foveating) and, more importantly, our attention towards a particular external object most definitely alters perceptual *representations* and perceptual *experiences* (Carrasco et al. 2004; Carrasco 2011). Similarly, the inner-sense theorist can comfortably accept that directing introspective attention to a conscious experience alters it.

Shoemaker (1988; 1996) criticism of what he calls "broad" perceptual-like mechanisms of introspection is worth considering. 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. Self-blindness, according to Shoemaker, is impossible. Then, the following principle emerges:

> SELF-TRANSPARENCY: 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.[13]

According to Shoemaker, inner-sense approaches to introspection (what he calls broad perceptual models) would deny SELF-TRANSPARENCY making self-blindness true; in particular, they make possible that the target conscious state exists independently from the subject being introspectively aware of it (e.g. if the detection mechanism were absent or inoperant): "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 don't hurt him. But this of course is nonsense." (Shoemaker 1994, Lecture II, p. 275)

Once again, I do not have the space to offer an in-depth analysis and rebuttal of Shoemaker's view. (Williamson (2002), Srinivasan (2015) and others have offered powerful demonstrations against arguments along the lines of SELF-TRANSPARENCY). Rather, my goal in bringing SELF-TRANSPARENCY up is to help contextualize some of the assumptions that iSDT makes. Shoemaker's rejection of self-blindness considers, but then rejects—in my view—too quickly, the possibility that introspection has a whole range of reliability. His famous example of being in pain and the impossibility of being introspectively unaware of such pains abstracts away from the important role the intensity of the pain plays (as well as other cognitive factors such as attention, cognitive load, etc.). In what follows, I first show that some kind of self-blindness is indeed possible and, second, that when the intensity of experiences is considered, the different accuracy ranges that introspection could have emerge in a way that Shoemaker's position does not really allow to.

[13] I borrow this reconstruction of self-blindness from Stoljar's (forthcoming).

iSDT assumes that consciousness is distinct from introspection, and hence that the introspective mechanism could break down without affecting conscious experiences *per se*. There is abundant evidence that first-order mental states can remain intact while introspective mechanisms break down. In cases of metacognitive failure, subjects display normal performance in perceptual or memory tasks while also displaying severe introspective limitations. These effects have been observed in both neuropsychological populations (Fleming, Ryu, Golfinos, & Blackmon, 2014) and causal interventions in neurotypical subjects (Cortese, Amano, Koizumi, Kawato, & Lau, 2016; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). While these effects do not show complete self-blindness (the lesions aren't absolute either), they are not subtle. Fleming et al. (2014) showed a 50% reduction in metacognitive efficiency in patients with prefrontal cortex lesions when evaluating their performance in a perceptual task in which they performed at normal levels.[14] Introspective mechanisms, then, can fail while keeping intact the conscious experiences that guide subjects during a task. This is evidence that consciousness and introspection can remain separate making (a kind of) self-blindness indeed possible.

In iSDT, weak pains and strong pains 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.

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 granularity with respect to the intensity of the mental state remains a problem. For example, if M is substituted for "a very weak pain" iSDT wouldn't predict that "most of the time" or even "ideally" you would believe that you're in (a weak) pain. In contrast, iSDT *would* predict that "most of the time" or "ideally" you acquire such a belief when the pain is strong. I take this to be an important prediction of iSDT that accommodates introspection as a garden variety mechanism in line with the rest of our cognitive capacities.

In the next sections I will offer the building blocks of iSDT's new 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

[14] Strictly speaking, a view where introspection and consciousness are not independent could still make use of the introspective signal detection paradigm I develop here. For example, someone who holds that all conscious states are introspected or that for a state to be conscious it is required that it is introspected, could still agree that the intensity of the conscious experience is linked to the accuracy of the introspective judgments. In the rest of the article I keep assuming that consciousness and introspection are distinct and so I don't consider this case further.

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 Primer

Consider the three following scenarios. All of them assume there is a man in the 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 notice the man next to the dumpster. His face looks bright and the contours of his facial features well-defined. You are confident that you saw someone.

DARK ALLEY: The alley's lamp is off, so the man's face looks dark and the contours of his face ill-defined. It is hard for you to see him and easy to take him for being just a shadow. You, mistakenly, categorize the alley as being empty. However, you are not confident about your decision.

DARK ALLEY + NEWS: Identical to DARK ALLEY except that you've heard that a robber was on the run in your neighborhood. You categorize the shadow as someone's face. Note that the shadow is visually processed in the same way it is processed in the DARK ALLEY scenario. The only difference is that knowing there is a robber on the run changes your perceptual judgment. You're still not highly confident about your perceptual decision.

These three scenarios illustrate three paradigmatic features of perception that SDT can successfully explain: perceptual sensitivity, response bias & confidence.

In a nutshell, according to SDT, perceptual judgments are determined by sensory sensitivity (i.e. the ability to distinguish between stimuli based on the way these stimuli 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, p. 20, see for a definition along these lines). In a paradigmatic case, perceiving is modeled as an observer deciding whether an internal perceptual response *e* (also called a perceptual sample or perceptual evidence) was generated by a stimulus class S1 (e.g. "stimulus absent" or "grating tilting left") or S2 (e.g. "stimulus present" or "grating tilting right"). The internal perceptual response is assumed to be a magnitude that corresponds to the strength of sensory evidence.[15] A fundamental assumption of SDT is that, across repeated presentations of the same stimulus, the internal perceptual response can be of different values due to ever-present random noise (either in the environment or in internal perceptual processing). The dimension along which the values of the internal perceptual response are distributed is called "the decision axis". Different stimulus classes generate a noise and a signal (plus noise) distribution of *e*, respectively.

3.1 Sensitivity

The noise and signal (plus noise) distributions have different means. For example, in a detection task (like the one taking place in the ALLEY scenarios), *e* represents the strength of the sensory evidence in favor of saying the stimulus was present (rather than absent) on a particular trial. The typical value of *e* would tend to be higher for stimulus-present trials than for stimulus-absent trials, which is reflected in the means of the distributions (Figure 1). The distance between the distributions' means (relative to their standard deviation) determines the observer's sensitivity (*d'*).[16] In other words, the signal-to-noise ratio of the distributions is the SDT measure of 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 1, top panel) than in the DARK ALLEY and the DARK ALLEY + NEWS scenarios (Figure 1, middle and bottom panels).

2.2 Response bias

The distributions always have some degree of overlap. Thus, it is always possible for a given value of *e* to have been generated by S1 or S2. Classifying *e* as S1 or S2 always involves uncertainty.[17] To make a perceptual judgment, observers classify *e* as S2 if it exceeds a response criterion *c* (solid vertical lines in Figure 1), and as S1 otherwise. Importantly, whereas

[15] The internal response is understood in SDT as a hidden psychological variable. At the implementation level, SDT can be further adapted to describe the neural underpinnings of internal responses (Shadlen & Kiani, 2013).

[16] This assumes the distributions are normal (Gaussian) and that they have equal variance.

[17] This uncertainty need not be reflected in the subject's subjective confidence. Subjects may feel confident about their perceptual decisions and yet the internal response evidence be ambiguous—or at least not certain—between pertaining to the S1 or S2 distributions. Whether this ambiguity also is reflected in subjects' phenomenology is more contentious. Recently, the question of whether there is perceptual confidence (i.e. phenomenology of confidence in perceptual experiences themselves), or more generally whether perception is probabilistic, has been widely discussed (Beck, 2019; Denison, 2016; Gross, 2020; Morrison, 2016; 2017; Munton, 2016; Siegel, n.d.). Here I'm neutral as to whether perceptual phenomenology reflects confidence or not. Confidence *judgments* may turn out to be based on perceptual confidence, but as I discuss below, the SDT apparatus doesn't require it.

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.).

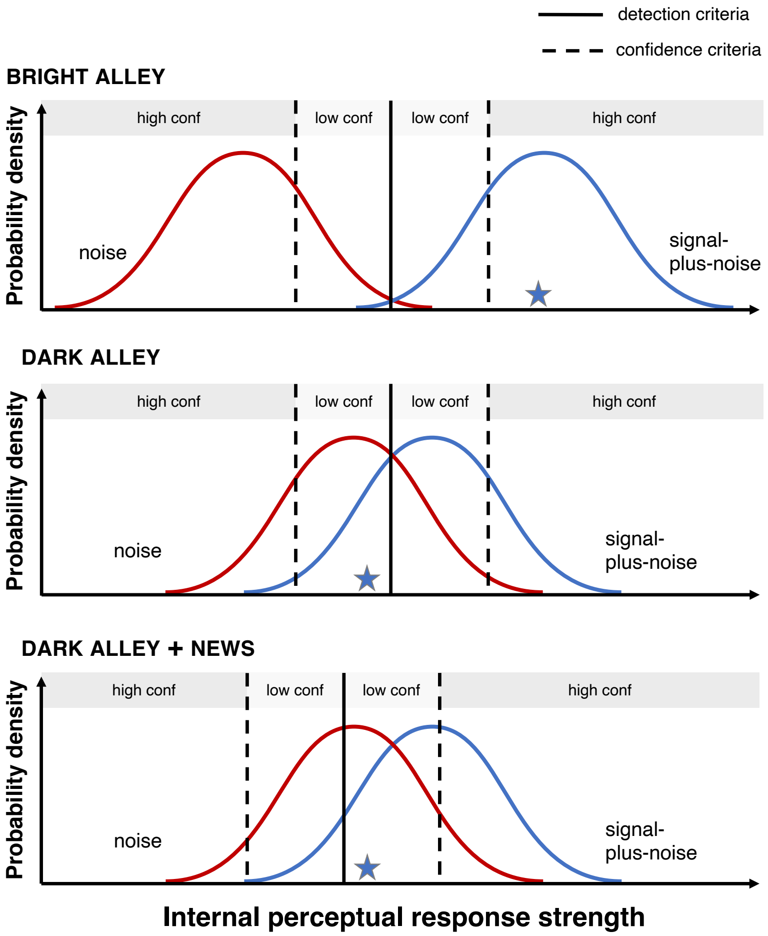

Figure 1. Dark alley scenarios modeled by SDT. The distance between the red (noise) and blue (signal-plus-noise) curves represents the observer's perceptual sensitivity (d'). In BRIGHT ALLEY (top panel), the internal perceptual response strength (blue star) produced by the stimulus (i.e. the man) crosses both the detection criterion (solid line) and the right confidence criterion (dashed line). The observer accurately detects the presence of the man with high confidence. In DARK ALLEY (middle panel), the observer inaccurately classifies the alley as being empty, with low confidence (the internal perceptual response only crosses the first confidence criterion but not the detection criterion). In DARK ALLEY + NEWS (bottom panel), the exact same internal response as in the DARK ALLEY scenario for an observer with the exact same sensitivity produces an accurate detection of the man's presence in the alley (with low confidence). In this scenario, the detection criterion is shifted to the left making the observer more liberal. For simplicity, the same is assumed to be true of the confidence criteria.

Importantly, as the DARK ALLEY and the DARK ALLEY + NEWS scenarios illustrate, sensitivity and response bias are independent from each other. While preserving identical sensitivity (i.e. the distance of the means of the distributions in both cases are the same), an identical internal perceptual response can yield different perceptual judgments due to a difference in the position along the decision axis where the criterion is placed. The criterion for detecting the presence of people becomes more liberal when the observer learns the news about the robber in the neighborhood. For an observer with identical sensitivity and different criteria, an identical internal response can yield a correct classification as in the DARK ALLEY + NEWS scenario (a hit in SDT terms) or an incorrect classification as in the DARK ALLEY scenario (a miss in SDT terms).[18]

3.3 Confidence

Ratings of confidence in one's perceptual judgments can also be characterized as resulting from a criterion-setting process (dashed lines in Figure 1). Confidence levels are determined by confidence criteria that further partition the decision space. When the internal response crosses both the detection criterion and the confidence criterion, observers report detecting the stimulus with high confidence (BRIGHT ALLEY scenario; Figure 1, top panel). If the internal perceptual response crosses the detection criterion but fails to cross the confidence criterion, observers report detecting the target but with low confidence (DARK ALLEY + NEWS scenario; Figure 1, bottom panel). An analogous explanation in the other direction applies too. Observers will report not detecting the stimulus with low confidence when the internal perceptual response crosses the left confidence criterion, but not the detection criterion (DARK ALLEY scenario; Figure 1, middle panel). Finally, when the internal perceptual response is weak enough to cross any criteria, observers will judge with high confidence that the stimulus is absent.

This brief introduction to signal detection theory highlights the crucial role the internal perceptual response plays. To successfully explain introspection based on insights from SDT, iSDT needs an equivalent notion: an internal *introspective* response. I offer a plausible candidate in the next section.

[18] Comparable accurate and inaccurate judgments can take place too when there is no stimulus (i.e. a "stimulus absent" trial). In SDT terms, in this case perceptually judging that there's no stimulus is a correct rejection; judging the stimulus is present is a false alarm.

4. Mental Strength & the 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. 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 internal perceptual responses and weak stimuli typically produce weak internal perceptual responses. As explained in the previous section, perceptual sensitivity is a function of the strength of internal perceptual responses over time.

To model introspection the way SDT models perceptual sensitivity, iSDT needs a functional analogue of internal perceptual responses. Since these are abstract postulates (hidden variables), iSDT may also postulate internal *introspective* responses as hidden variables that play an analogous role when modeling introspective sensitivity. But what, if anything, produces internal introspective responses? And what are these proportional to? (See Figure 2)

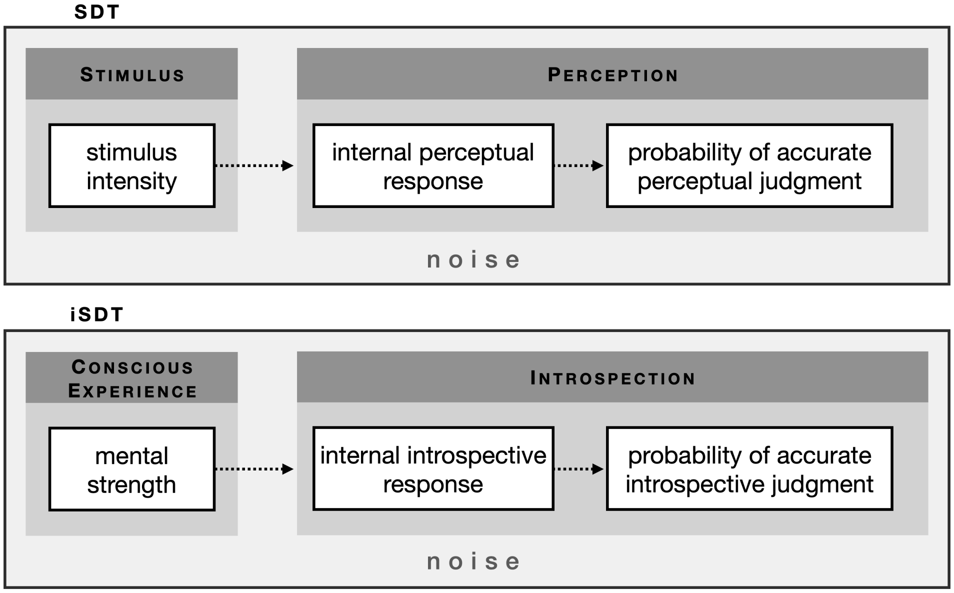
SDT
STIMULUS
PERCEPTION
stimulus intensity
internal perceptual response
probability of accurate perceptual judgment
noise
iSDT
CONSCIOUS EXPERIENCE
INTROSPECTION
mental strength
internal introspective response
probability of accurate introspective judgment
noise

Figure 2. SDT and iSDT models. According to SDT (top), stimulus intensity largely modulates (but not determines, as indicated by dotted arrows) the internal perceptual response in the perceiver. This, in turn, modulates the probability of making an accurate perceptual judgment. The more intense a stimulus, the larger the internal perceptual response and the larger the probability of making an accurate perceptual judgment. All these processes are non-deterministic as they are all corrupted by ever-present noise. iSDT (bottom) has an analogous structure. The intensity of a conscious experience (its mental strength) modulates (but doesn't determine) the internal introspective response, which in turn modulates the probability of making an accurate introspective judgment. As in perception, none of these processes involved in introspection are deterministic as they are all corrupted by ever-present noise.

According to iSDT, the strength or intensity of conscious experiences—"mental strength" for short—modulates the strength of internal introspective responses, which, in turn, modulates 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 striking. We can rely on this obvious fact about our conscious experiences to begin calibrating our introspective reliability.[19]

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. Or to put it slightly different, mental strength is the degree of prominence that a conscious experience has in one's phenomenal field at a given time (Hill, 1988). Thus, an intense pain "takes over" a larger portion of one's phenomenal field than a mild pain; a vivid mental image is more intense than a faint one; an experience as of a striking sound is stronger than the experience as of a loud sound.

Importantly, there are a few clarifications about the relationship between stimulus intensity and internal perceptual response on one hand, and mental strength and introspective response on the other. Often, strong stimuli produce experiences with high mental strength (and weak stimuli produce weak experiences). In the case of pain, under normal circumstances the larger the (potential) tissue damage is, the stronger the pain. In visual imagery, the clearer, sharper, more detailed the imagined objects are, the more intense the visual image is. The same applies for perception: the stronger the stimulus is, the stronger the perceptual experience.

This correlation, however, doesn't always hold. For example, in the case of phantom limb pains, there is no tissue to be (potentially) damaged; and yet, they can be intense. Under some conditions of high adrenaline, large tissue damage may produce little to no pain. Similarly, a vivid visual image of a very faint candle in a very dark room may in fact lack any clear details.

[19] One may wonder whether the obviousness of mental strength is something we know introspectively. If answered affirmatively, one may worry that the explanation of introspective accuracy based on mental strength is unwarranted if the variability of mental strength is based on introspective judgments (which might be inaccurate for all we know). I don't think we need to worry about this. Take perception. One may suffer from very inaccurate perception about a set of stimuli and yet be reliably aware that the stimuli are different along a particular dimension. Concretely, you may fail to perceive the correct orientation of a series of lines and yet accurately perceive that they differ in their orientations. The obviousness of the variety of degrees of intensity of conscious experiences works in a similar way. Even if introspection is fallible and you could be wrong determining accurately the intensity of any of your experiences, you may still reliably introspect that they have different intensities.

The intensity of experiences can be similarly decoupled from the strict intensity of the stimuli that produce them. For example, listening to loud music during a party may not be experienced as intensely as if the exact same sounds were played during night time in the tranquility of your home. Alternatively, experiencing *extreme* silence can produce intense *auditory* experiences. People pay good money to experience sensory deprivation tanks, in part, I surmise, to have an overall intense experience produced by minimal sensory stimulation. Finally, perception can take place unconsciously. At least in principle (but this might depend on one's views about consciousness), the strength of the internal perceptual response may dissociate from the intensity of a conscious experience. Highly accurate unconscious perception (e.g. blindsight) requires consistent strong internal perceptual responses that, however, do not lead to consciousness.

Similarly, the strengths of internal introspective responses typically (but not necessarily) will depend on the strength of conscious experiences. In normal cases, intense, vivid experiences produce strong introspective responses. But due to noise and other factors (e.g. attention, cognitive load, etc.) a weak experience might create a strong introspective response or a strong experience might create a weak introspective response. There is a close modulation of introspective responses by mental strength, but it is not a one-to-one relation. Moreover, as explained above, mental strength does not always depend on the strength of the stimuli and, in consequence, the strength of introspective responses does not always depend on the strength of the stimuli either.

The details of a theory of mental strength need not be spelled out here (see Morales "Mental Strength", *manuscript*, for details). 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. Introspective Signal Detection Theory

iSDT provides the tools for capturing sensitivity, response bias, and confidence of introspective judgments in a systematic manner. In what follows, I will focus mostly on cases of pain, but what I say here should be applicable to other conscious experiences as well.

First, let us characterize the phenomena we want the theory to capture. In the following scenarios the assumption is that you are in fact experiencing a pain you are introspecting.

STRONG PAIN: You wake up with a very strong back pain. It's excruciating. You head to the ER. A nurse asks if you're confident you are in pain. When you introspect your experience, you accurately judge that you're experiencing a strong back pain and you do it with high confidence. The nurse's question seems odd to you—of course you're confident you're in pain!

MILD PAIN: An hour after taking powerful painkillers, your back pain is mild. Mild enough that it's now hard to tell if you're still in pain. When you introspect your experience, you accurately judge that you have a mild pain but you do it with low confidence. When the nurse asks if you're sure you're still in pain, the question doesn't seem so odd anymore: you are actually not totally sure (or at least you are definitely *less* confident than when your pain was strong).

MILD PAIN + PRESENTATION: This scenario is identical to MILD BACK PAIN (i.e. by stipulation, the intensity of the pain is identical in both scenarios). Here, however, you have to give an important presentation later in the day; you really can't be in any kind of pain. When you introspect your mild pain, you honestly judge that you're not in pain anymore. Your confidence in your judgment is low (or, in any case, lower than when you pain was strong).

Admittedly, while STRONG PAIN—and to some extent MILD PAIN—are likely to be considered plausible by most, MILD PAIN + PRESENTATION may strike some as counterintuitive. The intuitiveness of the cases, however, cannot be judged introspectively under risk of getting it wrong (as discussed at the outset, this is 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. I suspect this explains in part the resistance to the idea that we can be introspectively wrong (or self-blind) about pains and other conscious experiences. Intuition, introspection, or correction by 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 possibilities and make these discoveries. By appealing to a similar kind of machinery to the one supporting SDT, iSDT can provide a systematic account of the three PAIN scenarios.

Similarly to how perceptual judgments are conceived in SDT, introspective judgments in iSDT are determined by sensitivity (i.e. the ability to distinguish between conscious experiences based on the way these experiences 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 as to whether an internal introspective response *i* (i.e. introspective evidence) was generated by a conscious experience of class C1 (e.g. "no pain" or "throbbing pain") or C2 (e.g. "pain" or "stinging pain"). The internal introspective response (or introspective response for short) is assumed to be a magnitude that corresponds to the strength of introspective evidence which, in turn, is correlated with the degree of conscious intensity of the experience (i.e. its mental strength). Repeated attempts to introspect similar experiences produce introspective responses with different values due to noise of different sorts. Introspective noise and an introspective signal distributions are generated by repeated encounters with C1 and C2, respectively. The values of the introspective response are distributed across a decision axis.

5.1 Introspective sensitivity

The distance between the distributions' means determines the introspector's sensitivity.[20] In STRONG BACK PAIN, the distributions don't overlap much, indicating high introspective sensitivity for strong back pains (Figure 3, top panel). The intense back pain produces a strong introspective response that is easy for you to introspectively classify as a strong pain (i.e. it's far from the detection criterion).

The situation for mild pains is quite different. In MILD BACK PAIN and MILD BACK PAIN + PRESENTATION, introspective sensitivity is lower because the distributions overlap more (Figure 3, middle and bottom panels). The mental strength of mild pains tends to be less intense which in turn tends to make the introspective response weaker. 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. The area under the noise curve (red) that crosses the detection criterion is larger in MILD BACK PAIN & MILD BACK PAIN + PRESENTATION than in STRONG BACK PAIN, which is iSDT's way of modeling the increased probability of false alarms. (An analogous explanation can be given for misses).

Note that the probability of error in STRONG PAIN is very small (i.e. the area of overlap between the two curves in the top panel of Figure 3). Introspective errors under these conditions should be quite rare (just as it's rare to fail to see a man in a bright alley who is in your line of sight). However, introspective errors under these optimal conditions are not impossible (just like perceptual errors in optimal conditions aren't impossible either). You could introspectively miss a very strong pain, for example, if the introspective response fails to cross the detection criterion (Figure 3, top panel). This could happen even if the mental strength of a particular painful experience is strong. For example, as the strong pain

[20] That is, with some background assumptions about the shape of the distributions.

experience is getting transformed into an introspective response, this could become weaker (i.e. weaker than the mental strength of that kind of pain would normally generate). In a case like this, you could be in strong pain and yet introspectively judge that you aren't (because the weakened introspective response fails to cross the introspective detection criterion).

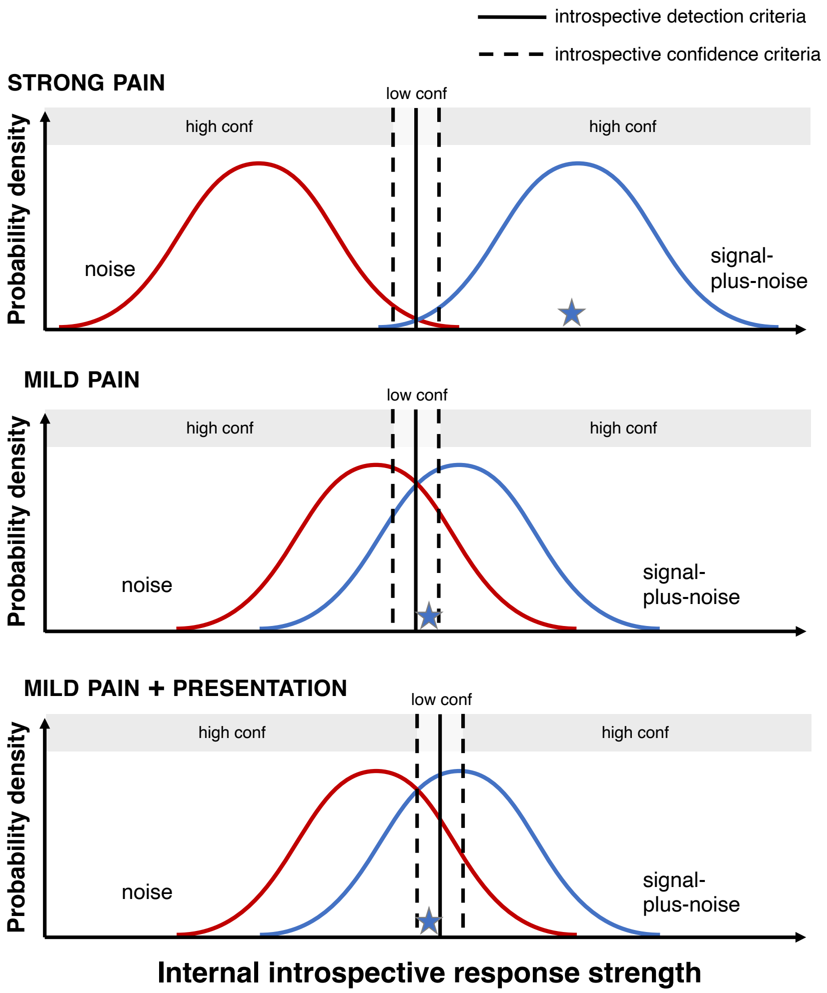

Figure 3. Pain scenarios as modeled by iSDT. iSDT models introspection as a function of sensitivity and response bias. In STRONG PAIN, a noise and signal-plus-noise distributions have little overlap and strong pains tend to produce strong internal introspective responses (signaled here by a blue star as if drawn from the signal-to-noise distribution). In MILD PAIN, the distributions overlap more and the introspective response is weaker, falling within the low introspective confidence partition but allowing the introspector to correctly introspect the presence of the pain. In MILD PAIN + PRESENTATION, everything is identical to MILD PAIN except that the introspector is more conservative. This means that their detection and confidence criteria are shifted to the right along the decision axis making the introspector to inaccurately introspectively judge that they are not in pain.

There can also be introspective false alarms (Figure 4, middle panel). For example, during so-called dental fear episodes, patients report feeling pain even before the dentist's instruments touch them (Meier et al., 2014; Rosenthal, 2005, p. 127). This can occur even in patients who have been anesthetized or whose tooth's nerves have been removed, making experiencing actual physical pain short from impossible. Rather, a way of explaining this phenomenon in a completely consistent way within the iSDT framework is to say that vibrations produced by the dentist's instruments in conjunction with the patient's fear of the treatment increases their introspective response which in turns produces a pain report even though no pain can be experienced under these circumstances.

5.2 Introspective response bias

Criterion effects can account for introspective variation too, even when holding introspective sensitivity fixed. In the MILD PAIN and MILD PAIN + PRESENTATION scenarios introspective sensitivity is, by stipulation, identical (Figure 3, middle and bottom panels). And yet in one scenario an identical pain is reported but it is not in the other. By shifting the introspective criterion, the degree of introspective response required for an introspective judgment of being in pain changes: in MILD PAIN + PRESENTATION external motivation makes your introspective criterion more conservative, preventing you from classifying the experience as of pain. Note that a case like dental fear could also be explained by a liberal criterion as opposed to by an increased introspective response, as I suggested above. The fearful patient is more willing to classify as pain a weak introspective response produced by no pain experience (Figure 4, bottom panel).

5.3 Introspective confidence

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 Figures 3 & 4). Note that to represent the common intuition that introspection is, if not infallible, at least unlikely to be significantly wrong in cases about pain (Alston, 1971; Dennett, 2002; Langland-Hassan, 2017; Rorty, 1970), the confidence criteria are placed much closer to the detection criterion. This means that the majority of introspective cases, no matter what, will be introspected with high confidence because the introspective response crosses the confidence criterion in most cases. iSDT leaves room for doubt or lesser confidence (those few instances where the introspective response falls between the detection and the confidence criteria), but it captures the fact that *in most cases* our introspective judgments are made with high confidence (pace Schwitzgebel 2008).

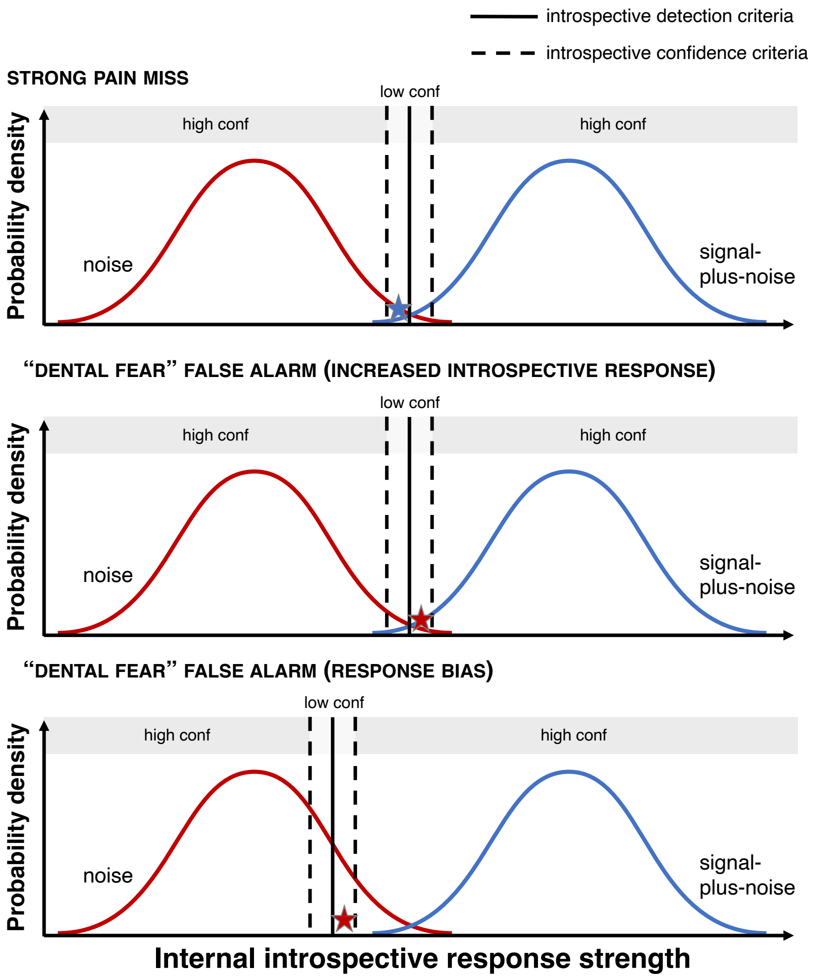

Figure 4. Inaccurate introspection scenarios. Even when introspecting a strong pain (or the lack thereof) introspectors may be inaccurate. In STRONG PAIN MISS, the transformation of the intense conscious experience into an introspective signal produces a rare weak introspective signal that does not cross the detection criterion. This results in an inaccurate introspective report that the introspector is not in pain. In DENTAL FEAR FALSE ALARM (INCREASED INTROSPECTIVE RESPONSE), fear of the dental procedure increases the introspective response of the non-painful experience (indicated by a red star to signify it should be classified as if drawn from the noise distribution). This makes this noise response to cross the detection criterion producing an introspective false alarm. In DENTAL FEAR FALSE ALARM (RESPONSE BIAS), the introspective signal is, as expected weak, but due to fear the detection criterion is shifted, making it more liberal. The introspector misclassifies the introspective response as coming from the signal (i.e. pain) distribution.

5.4 Beyond pains

Another advantage of iSDT is that it models introspection in the same way for any kind of conscious experience. All that is required is that the conscious experience has a degree of intensity (i.e. mental strength) that can create an introspective response. As discussed in section 4, not only pains but also mental imagery and perceptions have mental strength.[21] 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. Directing attention to it during introspection may increase its intensity even further (similar to Hill's (1988) "volume control hypothesis"). iSDT expects a case like this to produce a strong introspective response resulting in accurate, confident introspective judgments. But conjuring vivid mental images is hard. When these are weak or complex enough that simultaneously attending to all its features becomes hard, introspecting them is harder too. More inaccurate judgments with lower confidence are to be expected for the same reason mild pains were expected to produce this kind of judgments: lower sensitivity and weaker introspective responses that increase the probability of a misclassification and of a failure to cross the high confidence criterion.

Perceptual experiences should follow exactly the same mold. Strong experiences are more likely to be accurately introspected and with high confidence. As discussed in section 4, strong experiences need not be tied to strong stimuli. This leaves open the possibility to cases where a weak stimulus produces a strong experience and hence accurate, confident introspection. For example, a quiet but unexpected sound in the dead of night that attracts your attention; when introspecting this intense experience, you can confidently get it right.

iSDT explains "easy" cases at the same time that its machinery explains "hard" cases. For instance, the reason introspection is accurate in STRONG BACK PAIN explains, at least in part, the intuition that introspection might be infallible. After all, as discussed in section 2, the examples used to defend infallibility are often limited to simple cases of strong, intense experiences such as introspecting the experience of being hot, introspecting a strong pain, seeing a red tomato, etc. iSDT doesn't predict infallibility, but it does predict inaccuracy in those cases should be *very* unlikely. Importantly, by the same token, iSDT can explain the whole range of introspective reliability, including "less intuitive" cases of introspective inaccuracy and the conditions under which it might occur. Something similar can be said about response biases and confidence. Of more heuristic value than getting all the details right is that iSDT builds a *framework* to think about these and other cases in a systematic fashion.

[21] Other experiences also have degrees of intensity (e.g. itches, emotions, moods, action-awareness, sense of bodily ownership, etc.). The details for these cases shouldn't be different from the ones I discuss.

6. Introspection Is Not Perception

Introspection is signal detection. Introspection can be modeled after perception. But introspection is not perception any more than perception is receiving radio signals, the original domain of application for SDT models. Inner-sense models in general, and iSDT in particular, have been often unfairly attacked for allegedly committing to untenable similarities with perception. The similarities of introspection with perception, however, are limited to the type of inner mechanism that makes them possible. Just as we don't perceive our internal perceptual response (rather we classify it when we perceive), we don't perceive our internal introspective response (rather we classify it when we introspect). Here I want to discuss some important differences between the two.

The first difference is about the appearance/reality distinction. In perception, there is a clear distinction between reality and appearances. You can be wrong about seeing a light or hearing a sound. However, some deny that you can be wrong about whether there *seems* to be a light or whether there *seems* to be a sound. Recall Descartes's claim: "I certainly *seem* to see, to hear, and to be warmed. This cannot be false." Conscious experiences *are* the appearances, and, it is argued, there cannot be a further appearance/reality distinction involving these appearances. If there were one, there would be appearances of appearances. But an appearance of an appearance is indistinguishable from the first appearance.

If this is true of visual and auditory experiences, it is certainly true about pains.[22] Some may think that it follows that there is no difference between the appearance of being in pain and being in pain. According to this line of reasoning, appearing to be in pain should collapse into being in pain. But then we could not be introspectively wrong about being in pain. Even when you are not originally in pain, if it introspectively seems to you that you are in pain, your seeming to be in pain should collapse into actually being in pain. Contrary to one of iSDT's important tenets, this would make introspective failures impossible.

This might lead some to doubt that there is an appearance/reality distinction in introspection, and thus whether iSDT is tenable. But this rests on an ambiguity in the term 'seeming'. It is one thing to have an experience such that things "phenomenally seem" to you to be a certain way. It is another thing to make a judgment such that it "introspectively seems" to you that something is the case.[23] In both cases, things *seem* to you to be certain way, but they are very different in nature. Nothing about these two types of seeming necessitates a collapse

[22] It is a contentious issue whether pains can really be assimilated to perceptions. Hence, it is debatable what counts as the underlying reality of which pain experiences are appearances of (Aydede, 2009).

[23] See (Schwitzgebel, 2008) for discussion of this point.

of one into the other. Introspective judgments about pains (introspective seemings) do not collapse into painful experiences (phenomenal seemings). Thus, an appearance/reality distinction in introspection can be preserved.

This appearance/reality structure allows for introspective inaccuracies. According to iSDT, introspective seemings (appearances) need not match your phenomenal seemings (reality). Importantly, the existence of this difference has been recently confirmed by a linguistic analysis of internet searches according to which, English and German speaking users use "I feel pain" more often when describing minor or little pains, while they use "I have pain" significantly more often to describe severe or major pains (Reuter, 2011). Presumably, subjects follow the usage of feel/have in a similar way as in other modalities to express degrees of confidence indicative of an understanding of an appearance/reality distinction. Compare: "the shirt looks blue" vs "the shirt is blue," where the former is used to express low confidence about the real color of the shirt and the latter expresses confidence in the perceptual judgment. Similarly, subjects use "feel pain" for mild pains to express their weak confidence in their reported experience, while they use "have pain" to express certainty about the presence and characteristics of their experience. Thus, subjects indicate they are sensitive to when their phenomenal seemings (i.e. their experiences) match their introspective seemings and when they detect a possible mismatch.

The second difference between perception and introspection I want to discuss has to do with phenomenal character. In normal circumstances, conscious perception gives rise to a phenomenal character related to its object. For example, there is a special phenomenal character that accompanies perceiving an object as red. In contrast, introspecting a conscious experience arguably does not have a phenomenal character related to the target conscious experience. An introspective judgment of pain does not have a phenomenal character related to the pain or to painfulness. If it did, inaccurate introspection would give rise to incompatible phenomenal characters (e.g., simultaneous phenomenal characters of having a sharp and a dull pain in the same bodily location). But, to my knowledge, people do not report experiencing this kind of conflicting phenomenology.[24]

It is possible that introspective judgments have a *distinct* phenomenal character, as defenders of cognitive phenomenology would argue (Montague, 2015). In stark contrast to perception,

[24] Experiencing "impossible colors" is the closest case I can think of. Due to physiological constraints of our visual system, we do not perceive reddish-greens or yellowish-blues. In contrast, it is possible to see blueish greens or yellowish reds. However, under certain experimental conditions, some subjects report seeing an area in their visual fields as simultaneously green and red (Crane & Piantanida, 1983). Even if taken at face value, this result does not show that one can experience being in pain and not being in pain at the same time.

however, the new phenomenal character would be related to the act of judging, not to its target. Thus, introspecting a pain may have a particular phenomenal character provided it is not the phenomenal character of being in pain (for the reasons offered in the previous paragraph). Note that this is true too of perceptual judgments. In the Müller-Lyer illusion, you experience—it phenomenally seems—that one line is longer than the other. But you do not judge—it does not introspectively seem—that this is the case. This judgment, however, does not alter your *visual* phenomenology. You still experience the lines as having different lengths. Likewise, when you judge that you are in pain, this does not alter the phenomenology of the pain itself, even if there were a phenomenology of judging.

Let me finish this section by iterating that the differences between perception and introspection discussed in this section do not affect iSDT. The similarities between perception and introspection that I relied on to develop iSDT are related to the discrimination of signals, not to their outputs, their phenomenal characters, or their appearance/reality structures.

7. Conclusions

iSDT can explain why, sometimes, we can introspect accurately (in line with infallibilists); and it can also explain why, sometimes, we can expect to be inaccurate when we introspect (in line with skeptics). In other words, iSDT validates the intuitions of extreme, incompatible views. I take this to be a virtue of the proposal. Unlike other theories of introspection, including other inner-sense theories, iSDT 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. A mechanism that operates, at a fundamental level, in similar ways to other important 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.
Aydede, M. (2009). Is Feeling Pain the Perception of Something? *The Journal of Philosophy*, *106*(10), 531–567. http://doi.org/10.2307/20620203
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

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

Chalmers, D. J. (2010). The Character of Consciousness. New York: Oxford University Press.

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

Crane, H. D., & Piantanida, T. P. (1983). On seeing reddish green and yellowish blue. *Science*, *221*(4615), 1078–1080. http://doi.org/10.1126/science.221.4615.1078

Denison, R. N. (2016). Precision, not confidence, describes the uncertainty of perceptual experience. *Analytic Philosophy*, 1–22.

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

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

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.

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

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*(12), 151. 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?refreqid=search-gateway:9b0edec3d3651cb9e2e7e1170f3be50b

Montague, M. (2015). Cognitive phenomenology and conscious thought. *Phenomenology and the Cognitive Sciences*, *15*(2), 167–181. http://doi.org/10.1111/phc3.12053

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

Morrison, J. (2017). Perceptual Confidence and Categorization. *Analytic Philosophy*, 1–12.

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

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

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Picciuto, V. J., & Carruthers, P. (2014). Inner Sense. In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its Modalities* (pp. 277–296). New York: Oxford University Press.

Renero, A. (2019). Modes of Introspective Access: a Pluralist Approach, 1–22. 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.

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?refreqid=search-gateway:2601bb41a42df5c335d2f616d64a1c16

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.

Siegel, S. (forthcoming). How can perceptual experiences explain uncertainty? *Mind & Language*.

Smith, J. M., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18. http://doi.org/10.1038/246015a0

Spener, M. (2015). Calibrating Introspection. *Philosophical Issues*, *25*(1), 300–321. http://doi.org/10.1111/phis.12062

Srinivasan, A. (2015). Are We Luminous? *Philosophy and Phenomenological Research*, *90*(2), 294–319. http://doi.org/10.1111/phpr.12067

Stoljar, D. (Forthcoming). Armstrong's Just-so Story about Consciousness. In P. Anstey & D. Braddon-Mitchell (Eds.), *A Materialist Theory of the Mind Years On*.

Tanner, W. P., Jr. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

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.