# Internal inattentional blindness and the prospect of introspection training

Adam Morris

Harvard University

## 1 Abstract

Much of high-level cognition appears inaccessible to consciousness. Countless studies have revealed mental processes – like those underlying our choices, beliefs, judgments, intuitions, etc. – which people do not notice or report, and these findings have had a widespread influence on the theory and application of psychological science. However, the interpretation of these findings is uncertain. Making an analogy to perceptual consciousness research, I argue that much of the unconsciousness of high-level cognition is plausibly due to *internal inattentional blindness*: missing an otherwise consciously-accessible internal event because your attention was elsewhere. In other words, rather than being structurally unconscious, many higher mental processes might instead be "preconscious", and would become conscious if a person attended to them. I synthesize existing indirect evidence for this claim, argue that it is a foundational and largely untested assumption in many applied modalities (such as therapy and mindfulness practices), and suggest that, with careful experimentation, it could form the basis for a long-sought-after science of introspection training.

*Keywords:*

## 2 Introduction

Humans have been trying to perceive their own mental processes for a long time. Eastern contemplatives strived for millennia to become aware of what was happening inside themselves through meditative training (Shear & Jevning, 1999); in Western psychology, introspectionists like Wundt trained for thousands of hours to observe the workings of their own mind (Schwitzgebel, 2004), and early psychoanalysts like Anna Freud claimed that a central goal of therapy was to "bring into consciousness that which is unconscious" (quoted in Wilson (2004), p. 15). In contemporary society, widespread methods for mental improvement – such as cognitive therapies, mindfulness practices, and many popular personal/professional development techniques – purport to train awareness of the processes underlying one's judgments, choices, feelings, and so on (Dahl, Wilson-Mendenhall, & Davidson, 2020).

At the same time, modern experimental psychology has uncovered an enormous amount of high-level processes to which people do not seem to have introspective access. Starting with Nisbett and Wilson (1977), experiments have revealed fundamental unconscious processes underlying choice, judgment, attitudes, beliefs, and just about every other part of human experience (Bargh & Morsella, 2008; Evans, 2008; Haidt, 2001; Hassin, Uleman, & Bargh, 2004; Kahneman, 2011). These data are often interpreted as showing that people cannot reliably perceive the workings of their minds. For instance, in Wilson's (2004) influential book *Strangers to Ourselves*, he writes (p. 16):

> It is difficult to know ourselves because there is no direct access to the adaptive unconscious, no matter how hard we try. Because our minds have evolved to operate largely outside of consciousness, and nonconscious processing is part of the architecture of the brain, it may not be possible to gain direct access to nonconscious processes.

On this perspective, people can, at best, describe the current contents of their subjective experience or working memory (Dehaene, 2014; Ericsson & Simon, 1980), the

outputs of their mental processes (Nisbett & Wilson, 1977), or some abstracted features of those processes (like their overall accuracy; Fleming, Dolan, and Frith (2012)) – but, since so much of cognition happens unconsciously, most detailed reports of the processes underneath these surface-level experiences are untrustworthy (Wilson, 2004).[1] There is a tension between this view – that an enormous percentage of high-level cognition occurs unconsciously, and cannot be perceived introspectively – and the persistent societal belief that people can acquire substantial, accurate process-level self-awareness through training.

I argue that this tension results from a hypothesis which is confidently assumed by practitioners and rarely explored by experimental scientists: that some unconscious processes are unconscious only because of inattention. This hypothesis, which I call *internal inattentional blindness*, states that these unconscious processes are not *inherently* unconscious; their unconsciousness is not a structural or static feature of the mind. Rather, they are unconscious for the same reason that the gorilla in Simons & Chabris's (1999) classic study was unconscious: People fail to pay attention to them. If people did pay attention to the processes, they would become conscious – and hence people can learn, through internal attentional training, to directly perceive many of the actual processes underlying their experience and behavior.

In this paper, I lay out the internal inattentional blindness (IIB) hypothesis in detail and review existing evidence for it, of which there is much indirect and very little direct. I then sketch a roadmap for testing the hypothesis directly, and discuss its ramifications for both applied and basic psychology. The ramifications for applied psychology are clear: Since (as argued below) the hypothesis is widely assumed in popular modalities, testing it directly would either put these modalities on more solid scientific ground, or suggest that they need major revision.

The possibility of bringing unconscious processes into awareness also has significant implications for basic, experimental psychology. The scope of conscious

---

[1] Similar sentiments have been expressed in other influential work (Bargh & Williams, 2006; Carruthers, 2009; Gopnik, 1993; Greenwald & Banaji, 1995; Haidt, 2012; Johansson, Hall, Sikström, & Olsson, 2005; Kahneman, 2011; Wegner, 2017); see Section 7 for fuller discussion.

access is a fundamental question about the mind, and the conclusion that "there is no direct access to the unconscious, no matter how hard we try" (Nisbett & Wilson, 1977; Wilson, 2004) has not been seriously empirically challenged (Locke, 2009; Schooler, 2002a) – in part because there has not been a cognitively-grounded theory of how processes could transition from unconscious to conscious, or for how such claims could be tested objectively. IIB offers such a theory. If it is right, the IIB hypothesis would challenge Nisbett & Wilson's long-standing conclusion, and suggest that the observed ubiquity of unconscious processes in high-level cognition may not be a structural constraint of the mind, but rather an incidental reflection of internal inattention.

Perhaps most importantly, if the IIB hypothesis is right, it could offer a means of accelerating the pace of psychological science. Psychological science is, fundamentally, reverse engineering: Examining the mind from the outside, as one would an unfamiliar object (Pinker, 1997). While introspection should of course not replace objective data as the measuring stick of psychological truth, a training that verifiably improves people's ability to access their own mental processes could dramatically speed up this reverse engineering process – by guiding hypothesis generation, theory building, and experimental design down more profitable paths (Jack et al., 2003). Dehaene (2014) writes: "Obviously we cannot count on naive human subjects to tell us how their mind works; otherwise our science would be too easy". If trained subjects can actually, to some extent, tell us how their mind works, our science would plausibly get easier.

## 3 Conceptualizing internal inattentional blindness

### 3.1 Inattentional blindness towards external perceptual stimuli

Theories of inattentional blindness are built on two fundamental cognitive constructs: attention and conscious awareness. Though the two constructs are notoriously difficult to distinguish and define, attention has to do with the selection or prioritization of information for processing; when there are multiple things (e.g. visual stimuli) competing for limited-capacity processing, the mind prioritizes some at the expense of others, and the mechanisms that support this capacity are what constitute attention (Chun, Golomb, & Turk-Browne, 2011; Cohen, Cavanagh, Chun, &

Nakayama, 2012). In contrast, conscious awareness (in the sense I will focus on here) has to do with information becoming widely available across cognition, to the mind's "consumer systems" like verbal report, reasoning, rational control of action, and so on (Baars, 2002; Block, 1995, 2005; Dehaene & Naccache, 2001). Information that fails to enter consciousness (e.g. a subliminally-flashed word) can narrowly impact future cognition, through domain-specific peripheral processing or through effects like priming; but information that enters consciousness (e.g. the words you're reading right now) becomes globally available for report, inference, decision-making, recollection, and other high-level mental functions (Dehaene, 2014). This sense of conscious awareness has been called "access consciousness" (Baars, 2002; Block, 2005), and is associated with information being maintained in a long-range, brain-wide recurrent neural network (referred to as a "global neuronal workspace" (Dehaene, Changeux, & Naccache, 2011)).[2]

The precise relation between attention and awareness has been the subject of heated debate (Cohen et al., 2012; De Brigard & Prinz, 2010; Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Koch & Tsuchiya, 2007; Lamme, 2003; Van Boxtel, Tsuchiya, & Koch, 2010). But there is clear evidence for one strand of this relationship: Inattention can cause otherwise-reportable perceptual stimuli to fail to enter awareness. This is evidenced by three well-known patterns.

First, when people are engaged in an attentionally-demanding task, they fail to notice salient unexpected stimuli (Mack, Rock, et al., 1998). Famously, people fail to notice a gorilla walking unexpectedly through a visual scene if their attention is preoccupied (Drew, Võ, & Wolfe, 2013; Neisser & Becklen, 1975; Simons & Chabris, 1999). People also fail to notice a distinctive shape (e.g. a red square) unexpectedly

[2] There are several other relevant uses of the terms "consciousness" or "awareness". When a person is described as conscious of something (e.g. a visual stimulus), that statement is sometimes meant to indicate that (a) the person has qualia or phenomenal experience of the thing (Block, 1995; Nagel, 1974); or (b) they possess meta-representations about it – for instance, a representation of the veridicality of the original, "first-order" representation (like a confidence judgment of how accurately they saw a visual stimulus; Fleming et al. (2012); Lau and Rosenthal (2011)). Though both of these senses of consciousness are related to access-consciousness, they are conceptually separable from it (Dehaene, Lau, & Kouider, 2017), and the interrelationships are debated (Shea & Frith, 2019). Here, I will focus on access-consciousness and use the terms "consciousness" and "awareness" as such. But the metacognitive notion of consciousness is relevant to the discussion, and I will return to it later on.

appearing at fixation during a perceptual detection task (Mack et al., 1998; Most, Scholl, Clifford, & Simons, 2005); a tone being played in their ear while they're under high visual load (Macdonald & Lavie, 2011); or a clown unicycling across their path while they're talking on a cell phone (Hyman, Boss, Wise, McKenzie, & Caggiano, 2010). Of course, when people are paying attention, they easily notice all these stimuli. This phenomenon of inattentional blindness has been replicated and extended many times (Bredemeier & Simons, 2012; Memmert, 2006; Richards, Hannon, & Derakshan, 2010; Seegmiller, Watson, & Strayer, 2011; Simons & Jensen, 2009), and has been called "one of psychology's biggest exports" (Ward & Scholl, 2015).

Second, people fail to notice large, unattended-to changes in their visual field (known as "change blindness"; Simons and Rensink (2005)). People have severe difficulty identifying what changed between two sequentially-presented images, unless they are cued to attend to the specific locations where the changes occur (Rensink, 2002). In more dramatic demonstrations, people fail to notice when the sole actor in a movie is replaced by another person during an action scene (Levin & Simons, 1997), or when their conversation partner is swapped during a brief interaction (Simons & Levin, 1998).

Third, people exhibit an "attentional blink": After attending to one stimulus, they are briefly blind to a second stimulus (Raymond, Shapiro, & Arnell, 1992). In one demonstration of this phenomenon, people are shown a string of serially-presented letters ("C", "F", "X", etc.) and are told to note when any numbers (e.g. "7") appear. When two numbers appear a second or more apart, people easily notice and report both of them; but when the second number appears 200-500ms after the first, people who attended to the first often fail to notice the second (Shapiro, Raymond, & Arnell, 1997). This paradigm has also been replicated and extended many times (Dux & Marois, 2009), providing further evidence that otherwise-perceivable stimuli can go unnoticed when they are unattended to.

The fact that people routinely fail to perceive and report otherwise-accessible stimuli because they are not attending to them suggests that there are multiple reasons

a stimulus can be unconscious (Dehaene, 2014; Dehaene et al., 2006). Some stimuli are unconscious because of their intrinsic, structural features – e.g. a masked stimulus flashed for ten milliseconds is too weak to ever reach awareness. But other stimuli fail to reach awareness because of incidental features of the observer at the time of perception – namely, the observer was not attending to them. The former stimuli are in some sense permanently unconscious, but the latter could be brought into awareness with attentional control.

This observation is far from new, and features prominently in contemporary global workspace theories of consciousness (Fig. 1). In addition to distinguishing between conscious and subliminal (i.e. too weak to ever enter consciousness) stimuli, Dehaene et al. (2006) articulates an intermediate category: stimuli which were strong enough to enter consciousness, but did not because they were not attended to. Dehaene et al. refer to these stimuli as "preconscious" (Figure 1). The notion of preconsciousness plays a fundamental role in contemporary understandings of perceptual consciousness (Dehaene, 2014; Schooler, 2002b).

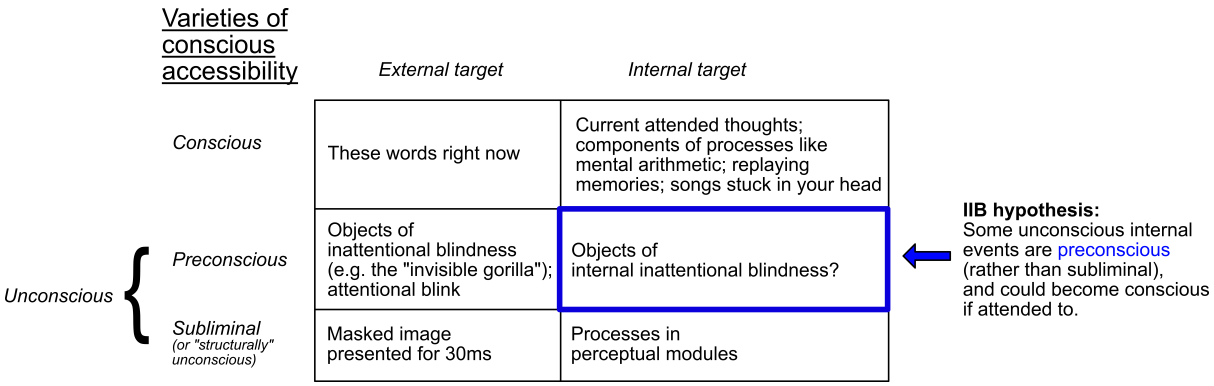


*Figure 1*. Varieties of conscious accessibility for external and internal targets. The tripartite distinction between conscious, preconscious, and subliminal was proposed by Dehaene et al. (2006), and some form of this taxonomy is now widely adopted in research on access consciousness (Dehaene, 2014). Though this distinction (and the concept of preconsciousness) has been applied almost exclusively to external, perceptual targets, it may apply to internal targets as well. The internal inattentional blindness (IIB) hypothesis is that some unconscious mental events (e.g. the activation of key representations during an unconscious decision process) are actually preconscious due to inattention, and could become conscious if attended to.

### 3.2 Inattentional blindness towards internal events

Inattentional blindness (and the existence of preconscious representations) has been studied primarily in the external domain: people failing to become aware of perceptual stimuli because their attention is elsewhere. It has not, however, been systematically applied to the domain of *internal* attention. As opposed to external attention, which operates over perceptual representations, internal attention is the selection or prioritization for processing of internal, cognitive representations: memories, task sets, goals, judgments, beliefs, decision options, plans, mental images, and so on (Chun et al., 2011; De Brigard, 2012; Dixon, Fox, & Christoff, 2014; Lückmann, Jacobs, & Sack, 2014). The boundaries of what exactly to characterize as internal attention are unclear, but plausible examples include selecting among competing task sets or goals (Chun et al., 2011; Gehring, Bryck, Jonides, Albin, & Badre, 2003); scrutinizing features of a mental image held in working memory (Fan & Turk-Browne, 2013; Griffin & Nobre, 2003; Souza & Oberauer, 2016); retrieving an item from long-term memory (Chun & Johnson, 2011; De Brigard, 2012; Logan, Cox, Annis, & Lindsey, 2021); and, more speculatively, bringing into awareness ongoing thoughts and cognitive operations (Fortney, 2020). Internal attention is similar to the executive controller posited in working memory models, although the precise relationship between these constructs is debated (Amir, Ruimi, & Bernstein, 2021; Awh, Vogel, & Oh, 2006; Chun, 2011; Kiyonaga & Egner, 2013; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012; Myers, Stokes, & Nobre, 2017). And though internal attention is dissociable from external attention, the two exhibit many similarities: they can overlap neurally (Kiyonaga & Egner, 2013), mutually interfere with each other (Kiyonaga & Egner, 2013), control which features of a representation get committed to memory (Fan & Turk-Browne, 2013), and be either intentionally controlled or "captured" by bottom-up salience (Van Ede, Board, & Nobre, 2020).

It is plausible, then, that just as inattentional blindness happens externally, it happens internally too; just as a person can fail to observe an unattended gorilla, she can also fail to observe internal events (like those preceding a judgment or decision)

because she was not attending to them. In other words, internal events could be preconscious: currently unconscious, but poised to enter awareness if attended to (Figure 1).[3]

What constitutes an "internal event"? The term is meant to capture the events which make up symbolic mental processes: the creation or activation of representations, and the operations which transform them into other representations. For instance, consider the mental process underlying the availability heuristic (Tversky & Kahneman, 1973). On this account, when a person judges the relative frequency of, e.g., homicides, they try to call to mind instances of homicide; they represent the number of examples they called to mind, or the ease with which they could generate examples (Schwarz et al., 1991; Schwarz & Vaughn, 2002); and they then combine that information with other background knowledge to arrive at a frequency estimate. For each of these component events – the activation of the example memories; the creation of the "ease of generation" representation; the activation of the other background knowledge; and the integration of those pieces of information into a final estimate – the person could, in principle, be conscious of them or not.

In practice, of course, people are often surprisingly unaware of the processes underlying their judgments, decisions, beliefs, attitudes, and so on (Nisbett & Wilson, 1977), and this fact has led many to the conclusion that we are "strangers to ourselves" (Wilson, 2004). The central claim of this paper is that some of those processes (or some component events of those processes) are plausibly like invisible gorillas – not structurally unconscious, just unattended to.[4]

---

[3] There is a trivial sense in which a great number of latent internal representations might be considered "preconscious": all the memories, plans, images, etc., which you currently aren't calling to mind but could at will. But this is a different sense of the term. Following Dehaene (2014), we restrict the label of preconsciousness to representations which are *currently activated* (not latent), and which are not currently in consciousness but would enter it if attended to. (In other words, preconscious representations must be represented "by firing", not "by wiring".) So your latent memory of today's breakfast is not preconscious; but if this memory gets activated while deciding what to eat for lunch, and (a) you don't notice the memory getting activated but (b) *would* have noticed if you attended to it, then that internal event was preconscious. (This discussion highlights the fact that there are more ways for representations to be unconscious than are relevant here; for a review, see Dehaene (2014).)

[4] Nisbett and Wilson (1977) famously argued that, though people typically have conscious access to the output of a mental process (e.g. the estimate of homicide frequency), they do not have access to the process itself. This process/output distinction generated significant controversy (Ericsson & Simon,

There are numerous theoretical nuances to this idea, as well as practical concerns about how it could be tested empirically. I will discuss these in Sections 5 and 6. First, however, I will give reasons to believe the hypothesis might be true.

## 4 Circumstantial evidence for the IIB hypothesis

There is much circumstantial evidence suggesting that some unconscious mental processes can be brought into awareness via internal attention. The largest source of circumstantial evidence comes from applied psychological modalities like therapy and mindfulness/meditation practices. In many of these practices, IIB is a foundational assumption, often so assumed that it is not explicitly named.

For instance, consider cognitive-behavioral therapy (CBT). CBT is a family of interventions considered the gold standard for evidence-based therapy (Hayes & Hofmann, 2017), and one of its focal points is helping people identify and reshape maladaptive thought patterns or schemas (Beck, 1979; Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012). A foundational assumption in CBT is that automatic thoughts (like the activation of negative schemas) play a core role in the mental processes underlying emotional reactions, choices, judgments, etc; that these thoughts are normally unattended to and unconscious; but that they can be brought into consciousness by attending to them, through conversation with a therapist and through practices like thought journaling or "attentional control training" (Beck, 1991; Beck & Haigh, 2014; DeRubeis, Webb, Tang, & Beck, 2010). Beck (1979), in a foundational text on CBT, wrote:

> Patients experienced specific types of thoughts of which they were only dimly aware and that they did not report... Unless they were directed to focus their attention on these thoughts, they were not likely to be very

---

1984; Smith & Miller, 1978; White, 1980). I only note here that there is no *a priori* reason why people can't be aware of the component events of processes as they unfold in real-time. Symbolic cognitive processes are a series of representations and computations – representations being activated and combined/transformed into different representations. People are clearly sometimes aware of activated representations (e.g. memories, plans, mental images, fantasies, worries, songs stuck in your head) and operations over those representations (e.g. mental rotation, mental arithmetic, rumination). It is an empirical question which processes, and which component events, people can attend to and become aware of.

> aware of them. Although these thoughts seemed to be on the periphery of the patients' stream of consciousness, they appeared to play an important role in the psychic life of these patients... It seemed to me that I had tapped another level of consciousness in the recognition of automatic thoughts, perhaps analogous to the phenomenon described by Freud as "preconscious"...

Similar observations can be found in mindfulness-based cognitive therapy (Segal, Williams, & Teasdale, 2018), dialectical behavioral therapy (Linehan, 2018), and many other modern therapies (Castonguay & Hill, 2007; Frank & Frank, 1993; Grant, Franklin, & Langford, 2002; Grosse Holtforth et al., 2007; Stein & Grant, 2014; Timulak & McElvaney, 2013). Pull out a modern therapy book, and it will likely be replete with examples of coming to notice important internal events that were going unnoticed before. The idea is so commonplace that it is rarely even presented as an important claim. Of course, not all approaches to therapy promote this kind of self-awareness (Wampold, Imel, Bhati, & Johnson-Jennings, 2007), and therapies differ enormously in what they say should be done with the thoughts once they are noticed; but for many modern therapies, overcoming IIB is a critical component.

IIB is also a foundational assumption in many mindfulness and meditation practices. For instance, meditation teacher Joseph Goldstein writes (Goldstein, 2017):

> What happens as the mind becomes silent and we become more finely aware, is that many of the things which were below our normal threshold of awareness, much of what is called subconscious material, become illuminated by mindfulness.

Similarly, Kabat-Zinn (1994), who popularized the term "mindfulness", argued that inattention leads to "a lack of awareness and understanding of our own mind and how it influences our perceptions and our actions"; that mindfulness practice "literally allows us to see more clearly" into our minds; that there is an "incessant stream of thoughts flowing through our minds" which we are typically unaware of, and that by

"focusing in on what the mind is up to", people can perceive these internal events and their influence on behavior.

Scientific treatments of mindfulness make similar claims. In their seminal paper, Brown and Ryan (2003) characterize mindfulness as "'the clear and single-minded awareness of what actually happens to us and in us'... Rather than generating mental accounts about the self, mindfulness 'offer[s] a bare display of what is taking place"' (Shear & Jevning, 1999; (Thera), 1972). In a similarly influential paper, Bishop et al. (2004) argue that a core component of mindfulness is "self-regulation of attention so that it is maintained on immediate experience, thereby allowing for increased recognition of mental events in the present moment... Rather than getting caught up in ruminative, elaborative thought streams *about* one's experience... mindfulness involves a direct *experience* of events in the mind and body." These sentiments are echoed in other theoretical accounts of mindfulness (Hadash & Bernstein, 2019; Lutz, Jha, Dunne, & Saron, 2015; Vago & David, 2012). Moreover, self-report measures of mindfulness ability typically include an assessment of how well someone can "observe, notice, or attend to a variety of stimuli, including internal phenomena, such as bodily sensations, cognitions, and emotions...". (Baer, Smith, & Allen, 2004; Baer et al., 2008; Young, 2016), and the practices often involve focusing attention on internal events (like thoughts, mental images, etc) for long periods of time (Goldstein, 2017; Young, 2016).

Moreover, this belief in some form of IIB extends outside formal therapy or mindfulness practice. It appears routinely, for instance, in coaching and professional development frameworks, like Google's popular "Search Inside Yourself" program (whose curriculum includes self-awareness training meant to "enhance your perception of your own emotions" and help you "accurately assess your thoughts"; Caporale-Berkowitz et al. (2021)), or the influential "immunity to change" coaching framework which relies on bringing unconscious change-resistant beliefs into consciousness (Kegan, Kegan, & Lahey, 2009). Indeed, a recent review of approaches to improving mental well-being cites acquiring "an experiential understanding of one's own psychological processes and how the dynamic interplay of these processes influences

experience" as one of the basic pillars of mental training (Dahl et al., 2020).

Finally, practitioners of these methods routinely report success in this endeavour. Both therapy clients and mindfulness trainees report improved awareness of their mental processes (Castonguay & Hill, 2007; Goldstein, 2017; Kabat-Zinn, 1994; Young, 2016), and this (self-reported) awareness mediates the other beneficial effects of training (Castonguay & Hill, 2007; Ghasemipour, Robinson, & Ghorbani, 2013; Hanley & Garland, 2017; Nakajima, Takano, & Tanno, 2019; Nyklíček, Zonneveld, & Denollet, 2020). For instance, patients consistently rate increased self-insight as one of the most helpful components of therapy (Castonguay & Hill, 2007; Hill & Knox, 2008); and people who practice mindfulness rate themselves higher on measures of internal awareness (e.g. "I have a very clear idea about why I've behaved in a certain way" (Nakajima, Takano, & Tanno, 2017) or "I don't know what's going on inside me", reverse scored (Nyklíček & Denollet, 2009)), with this reported awareness substantially mediating other positive effects (like reduced stress and depression, and improved quality of life) (Nakajima et al., 2019; Nyklíček et al., 2020).[5]

Of course, these modalities claim to involve much more than just improving awareness of internal processes. And there are certainly approaches to mental improvement which do not emphasize internal awareness; for instance, some clinical researchers argue that therapy gives people useful (but not necessarily veridical) beliefs or narratives about their mental processes without granting actual conscious access to those processes (McAdams, 1993; Wampold et al., 2007). Nonetheless, becoming conscious of previously-unconscious mental events is an essential component of a diverse, widespread set of successful applied modalities, and some form of IIB is a foundational theoretical assumption in many approaches to mental improvement (Dahl et al., 2020).

---

[5] These purported gains in internal awareness plausibly come from improved attention. Meditation training is known to improve attentional capabilities (Jha, Krompinger, & Baime, 2007; Lutz, Slagter, Dunne, & Davidson, 2008; Tang et al., 2007), leading to objective improvements in perceptual discrimination (Chan & Woollacott, 2007; Fox et al., 2012; MacLean et al., 2010; Slagter et al., 2007; Treves, Tello, Davidson, & Goldberg, 2019). Notably, practitioners experience at least some of these gains as coming from increased awareness of internal processes (Bernstein & Zvielli, 2014); for instance, Brown, Forte, and Dysart (1984) write that "phenomenological reports indicate that mindfulness practice enables practitioners to become aware of some of the usually preattentive processes involved in visual detection."

***Direct evidence?*** Of course, aggregated anecdotes do not sum to good evidence. For such a widespread and foundational premise, IIB has received strikingly little direct empirical investigation. Its plausibility is further suggested, however, by five lines of empirical work.

One is on people's awareness of their implicit attitudes. Hahn, Judd, Hirsh, and Blair (2014) found that, contrary to how implicit attitudes are commonly conceptualized, people appear quite aware of their implicit attitudes: Once cued to attend to them, they can report them with a high degree of accuracy (as measured by subsequent performance on an implicit association test). Moreover, when people are cued to attend to them, they seem to be recognizing something new in themselves; the experience of attending to and reporting them increases people's explicit acknowledgements of bias, and leads to higher explicit-implicit attitude correlations (Hahn & Gawronski, 2019). Thus, implicit attitudes seem like a candidate for a mental feature subject to IIB, which can be brought into awareness via attention. (This possibility is bolstered by the fact that people with mindfulness/meditation training show less of a divergence between their implicit and explicit attitudes (Carlson, 2013; Koole, Govorun, Cheng, & Gallucci, 2009; Remmers et al., 2018; Strick & Papies, 2017), and that people, when induced to attend to themselves, report attitudes more consistent with their subsequent behavior (Gibbons, 1983).)

A second promising line of research comes from Carpenter et al. (2019), who tested whether people can improve at their ability to recognize the accuracy of their own judgments. Participants performed a perceptual discrimination task, gave a "metacognitive" judgment of their accuracy on each trial ("how confident are you that you got it right?"), and then received real-time feedback on whether their metacognitive judgment was correct or not ("did you actually get it right?") (Fleming, Weil, Nagy, Dolan, & Rees, 2010). This feedback improved the accuracy of their metacognitive accuracy in both the perceptual discrimination task and a subsequent memory recognition task, suggesting domain-general improvement in metacognitive judgment through training. This improvement could be the result of increased conscious access to

the processes underlying perceptual discrimination or memory recognition. (It's also possible, however, that the training improved people's ability to infer their accuracy based on some other mechanism, like self-observation or inference; Carpenter et al. do not provide direct evidence for the mechanism of improvement.)

A third line of research is on the Libet paradigm (Libet, 1985). In this paradigm, people are asked to spontaneously perform an action and note the moment the intention to act forms in their mind. People report becoming aware of the intention several hundreds of milliseconds after its formation can be detected neurally (in the form of an EEG signal called the "readiness potential"), suggesting that the intention exists unconsciously before it becomes conscious. The interpretation of the Libet paradigm has been highly controversial (Maoz et al., 2015). Nonetheless, experienced meditators report becoming aware of the intention over 75 milliseconds before non-meditators do (Lush, Naish, & Dienes, 2016), and their reports show stronger correspondence with the timing and form of the readiness potential (Jo, Hinterberger, Wittmann, & Schmidt, 2015; Jo, Wittmann, Borghardt, Hinterberger, & Schmidt, 2014). These results could be explained if, consistent with the participants' self-report, the meditators had access to previously-unconscious internal events associated with intention formation.

A fourth line of research is on people's awareness of mental imagery. People can reliably report the strength of their mental images, as measured by the images' propensity to bias subsequent binocular rivalry displays (Pearson, Rademaker, & Tong, 2011). And the accuracy of those reports can be improved: People's judgments of their mental imagery strength become significantly more accurate with training (without a change in the actual strength of the images; Rademaker and Pearson (2012)). This result could be interpreted as people gaining increased conscious access to the activated mental image through training.

A final line of research comes from attempts to train internal awareness via direct coaching or guided interviews (Hurlburt & Heavey, 2001; Petitmengin, 2006). In these procedures, trained coaches guide participants through reporting their internal experience, helping them avoid pitfalls and focusing their attention directly on their

mental events. Examples include the "elicitation interview" method from the neurophenomenological tradition (Petitmengin, 2006; Varela, 1996), and the "descriptive experience sampling" method (Hurlburt & Schwitzgebel, 2011; Hurlburt & Heavey, 2001). These procedures provide suggestive evidence for the possibility of bringing unconscious internal experiences into awareness. For instance, epileptics who underwent guided interviews reported that they could later recognize and predict upcoming seizures (Petitmengin, Baulac, & Navarro, 2006); and mental patients who reported unusual internal experiences also exhibited corresponding behaviors (e.g. a woman who, in these interviews, described seeing many simultaneous images in her head also routinely watched three television screens at the same time; Hurlburt and Schwitzgebel (2011); Hurlburt and Akhter (2006)). Practitioners from both methods report many more compelling anecdotes (Hurlburt & Schwitzgebel, 2011; Petitmengin, 2009).[6]

But empirical studies on these methods have rarely used objective measures. A few studies found that, when people undergo these guided sessions, their reports of inner phenomena (like inner speech) correlate with sensible neural activity (Hurlburt, Alderson-Day, Kühn, & Fernyhough, 2016; Kühn, Fernyhough, Alderson-Day, & Hurlburt, 2014; Lutz, Lachaux, Martinerie, & Varela, 2002). The strongest experimental result comes from a study on "choice blindness" – the finding that people, when asked to explain why they made a choice which they didn't actually make, rarely notice the discrepancy and instead confabulate explanations for the non-existent choice (Johansson et al., 2005). Petitmengin et al. (2013) found that, after undergoing a guided interview about their internal experience, the percentage of people exhibiting choice blindness dropped from 77% to 20%. These results, however, are difficult to interpret due to major confounds in the design (Froese, Gould, & Seth, 2011).[7]

In sum, the consistent reports of practitioners, along with these many preliminary

---

[6] For example interviews, see Hurlburt and Schwitzgebel (2011) or the Supplementary Materials of Petitmengin, Remillieux, Cahour, and Carter-Thomas (2013).

[7] Interviews were performed immediately after people made their choice, and lasted between 15-45 minutes. On these interview trials, choice blindness was measured after this long interview process; on the control trials, it was tested immediately after making the choice. A much tighter control is needed to show that the reduction in choice blindness was actually due to improved awareness of decision processes.

experimental results, all suggest that the IIB hypothesis is plausible: that some mental events are unconscious solely due to inattention, and can be brought into awareness if attended to. But this evidence is indirect, and could all be explained in other ways. As a hypothesis with foundational implications for both the theory and application of psychology, IIB demands more systematic empirical treatment.

This lack of empirical investigation into IIB may be due, in part, to the difficulties of conceptualizing and testing for it. In the next two sections, I address some of these conceptual difficulties and then illustrate how the IIB hypothesis could be tested rigorously.

## 5 Theoretical concerns for the IIB hypothesis

I will address some theoretical concerns for the IIB hypothesis, and then in the next section illustrate how it could be tested.

***What is the scope of this hypothesis? Are you claiming that someone could become conscious of low-level perceptual processes?*** No. It is extremely unlikely that someone could ever become conscious of low-level perceptual processes, like the computations underlying depth perception or the representations activated in V1. These processes are almost certainly unconscious for permanent, structural reasons.

But for much of unconscious cognition, it is an open question which processes are structurally unconscious, and which are instead preconscious and could be brought into awareness if attended to. Intuitively, the likeliest candidates for IIB are high-level processes: processes underlying judgment and decision-making, belief and attitude formation, social psychology, etc. But the scope of IIB is a question to be determined empirically.

***There seem to be key aspects of mental processes that people cannot, in principle, be conscious of. For instance, even if people are conscious of performing mental rotation, they cannot report the computations that enabled that rotation (e.g. was it an affine transformation? which rotation algorithm was used?). How does this square with your theory?*** Being conscious of a mental process is not all-or-nothing. A person using the availability

heuristic to judge the frequency of homicides, for example, might notice that she is internally generating examples of homicides, but fail to notice that she is representing the ease of generating those examples. Or she might be aware of each high-level component event of the process, but not be able to describe them in low-level computational detail (i.e. she would likely not be aware of the mechanisms supporting her memory retrieval, or the precise algorithm she's using to combine the "ease of example generation" representation with other background knowledge about homicide frequency).

Obviously, there are levels of description of a mental process which people will never be conscious of. People will never be able to introspectively report the neural architecture supporting a mental process, and can likely never become conscious of low-level algorithmic details (like the cognitive architecture supporting memory retrieval).

These *a priori* limitations, however, do not in principle prevent people from gaining substantial conscious access to many high-level processes. For instance, consider a person who is initially unaware of how they're making frequency judgments, and then later reports that they are estimating frequency by observing the ease of generating examples. That person has substantially improved their awareness of the process – enough to describe the process in the same level of detail that it was originally described by researchers (Tversky & Kahneman, 1973). For many high-level mental processes, substantial process-level awareness seems possible without low-level knowledge of algorithmic details.

Moreover, if the IIB hypothesis is right, then it is an open empirical question how "deep" within a process conscious access can get. For instance, consider the process underlying multi-attribute choice – e.g. choosing between cars which vary simultaneously on price, horsepower, style, etc. In some contexts, people appear to weigh and combine the attributes linearly (Bhatia & Stewart, 2018); in other contexts, people appear to use nonlinear heuristics (such as choosing based on a single attribute (Gigerenzer & Todd, 1999), or eliminating items that don't have certain attributes

(Tversky, 1972)). Making this distinction involves detailed understanding of the underlying decision algorithm. And yet, if the IIB hypothesis is right, it seems plausible that people could become conscious of this distinction; someone could accurately observe, for instance, that they weighed together seven different attributes when selecting a car, but then chose solely based on price when selecting a beer. Again, the scope of internal attention and conscious access is something that needs to be determined empirically.

***What is the relationship between being conscious (as you're using the term here) of an internal event, and having meta-representations or meta-cognition about the event? Isn't meta-cognition crucial for introspective reporting?*** There are two main cognitive notions of what it means to be conscious of some information (like the occurrence of an internal event; Dehaene et al. (2017)). One is for the information to be widely available across cognition for processes like inference, rational control of action, planning, memory, etc. (Baars, 2002; Block, 1995; Dehaene & Naccache, 2001); as described above, this is often called access-consciousness, and is associated with the information being held in a "global neuronal workspace" (Dehaene et al., 2011). This is the sense in which we've been using the term "conscious" so far.

But another notion of what it means to be conscious of some information is to possess meta-representations about the information: e.g. to represent that you know the information, or to form a "second-order" representation about how accurate the "first-order" representation is (Fleming et al., 2012; Lau & Rosenthal, 2011). Common meta-cognitive assays, for instance, ask people to make perceptual judgments ("are these two low-contrast Gabor patches identical?") and then make meta-judgments about how accurate their first-order judgment was ("how confident are you that you were right?"; Fleming and Lau (2014)).

The meta-cognitive notion of consciousness is conceptually separable from access consciousness (Dehaene et al., 2017; Fleming et al., 2012; Shea & Frith, 2019). In principle, information ("those Gabor patches are identical") could enter a person's

global workspace – and be used for inference, planning, judgment, recall, and so on – without the person ever forming higher-order representations about the first-order information. And people can sometimes have higher-order representations without first-order awareness (e.g. in subliminal error detection; Charles, Van Opstal, Marti, and Dehaene (2013)).

Another way to conceptualize the difference is that being conscious of an event, in the global availability sense, is more aligned with *directly experiencing* the event. Of course, all cognitive theories of phenomenal experience are controversial, and some believe that phenomenology can occur in the absence of global availability (Block, 2011; Bronfman, Brezis, Jacobson, & Usher, 2014). Nonetheless, global availability and direct phenomenal experience seem deeply related (Dehaene, 2014). Meta-cognition, in contrast, is less conceptually associated with directly experiencing an event, and more associated with *thinking about* the direct experience of the event (Fleming et al. (2012), but see Lau and Rosenthal (2011) for an alternate view).

These two conceptions of being conscious of something, then, lead to different notions of what it means to have introspective access to an internal event or a mental process (Dunne, Thompson, & Schooler, 2019; Lutz et al., 2015). On the meta-cognitive sense, having introspective access to a mental process would mean having higher-order representations of the mental processes (e.g. knowledge that you make frequency judgments by generating examples). On the global availability sense, it would mean directly observing those internal events (e.g. example generation) as they're happening, and those events being subsequently available for report, inference, recollection, and so on. It is the latter sense of introspective access that I'm concerned with, and that internal attention may be able to facilitate.

Of course, meta-cognitive reasoning is still a crucial part of self-knowledge and of what it means to introspect (Fleming et al., 2010). But it would not be news to claim that people can acquire meta-cognitive beliefs about unconscious processes – that happens every time you read a convincing psychology paper. The novel possibility suggested by IIB is that internal attention can bring unconscious processes into direct,

access-conscious awareness (Dunne et al., 2019; Lutz et al., 2015).[8]

***If the IIB hypothesis is true, shouldn't it be obviously verifiable? Why can't I just turn my attention towards my mental processes and easily report them, like I can with the invisible gorilla? Why don't people start accurately reporting on their mental processes as soon as you alert them to their errors?*** Even if bringing unconscious mental processes into awareness is possible, it is plausibly quite difficult. It is the rare therapist or mindfulness teacher who would instruct a client to pay attention to an internal process and expect immediate success. Rather, attending to internal processes is supposed to be a skill that takes time and effort to develop.

The difficulty of attending to mental events is vividly illustrated by the "guided interview" methods described above (Hurlburt & Akhter, 2006; Petitmengin, 2006). In these methods, a practitioner leads a subject through an interview about their experience at a moment in time, helping them attend carefully to their internal processes. It is a laborious process, often taking upwards of half an hour for the subject to describe their internal experience at one moment in time. And subjects are, at first, really bad at it; it is only with much coaching that subjects begin to report anything plausibly resembling their actual experience.

Why is it so hard? Anecdotally, when asked to introspect, people habitually do other things instead; they "flee from actual phenomena and distort or mask them in a variety of ways" (Hurlburt & Akhter, 2006). For instance, they do the things Nisbett and Wilson (1977) describe: They fall back on *a priori* theories or shared cultural suppositions, or make inferences about themselves. As Petitmengen puts it: "When asked to describe a given cognitive process, our natural tendency is to slip surreptitiously from the description of our actual experience toward the verbalization of justifications, beliefs, explanations, generalizations, and abstract knowledge about our experience" (Petitmengin et al., 2013). Accurately introspecting on mental processes

---

[8] Even if the two types of introspection are conceptually distinct, they are in practice entangled; meta-cognition surely depends in part on direct experience of internal events (Morales, 2021), and having direct experience of mental processes would surely facilitate meta-cognition about them.

requires learning to carefully sort through these layers of distortion.

Moreover, internal attention can be difficult to control with precision. The mental processes underlying judgments, choices, etc., happen extremely quickly, on the order of tenths or hundredths of a second; people do not have practice internally attending to events that fleeting. This problem is exacerbated by the fact that internal events are likely much weaker than typical perceptual targets of external attention (e.g. "seeing" a mental image is typically harder than seeing an actual image). Yet another obstacle is that, for people to focus on the mental processes underlying their behavior, they must attend internally to those processes while simultaneously attending to whatever they are doing externally – thereby splitting their attention and taxing cognitive resources even further (Lutz et al., 2015). (Dividing attention between external tasks is hard, and often requires intensive training; people take upwards of 30 hours of practice, for instance, to become proficient at reading stories while copying words being spoken aloud (Hirst, Spelke, Reaves, Caharack, & Neisser, 1980; Spelke, Hirst, & Neisser, 1976). It is likely similarly difficult to divide attention between external and internal targets.) It is unsurprising, then, that practices like mindfulness or cognitive therapy so often involve building attentional capacity (Lutz et al., 2008; Segal et al., 2018).

A final reason that IIB may be difficult to overcome is motivational: Uncovering unconscious processes can be an upsetting and identity-shaking experience, as the theories or rationalizations we have about ourselves often diverge from how we actually operate. People may get glimpses of their hidden processes and then quickly flinch away (Petitmengin et al., 2013).

Much of this is anecdeotal, and precise answers may have to await a more developed account of internal attention training and IIB. But these reports suggest that simply instructing people to attend to a target mental process won't be enough. People need training.

## 6 How to test the IIB hypothesis

Despite its potential importance (and its widespread assumption in applied modalities), the possibility of IIB – and of bringing unconscious mental processes into

awareness – has received very little direct empirical investigation with objective measurements. This lacuna is likely in part due to the perceived difficulty of testing for it. Consider any mental process which research has discovered, but which people report no awareness of (or report inaccurately on). How in principle could we determine whether this process is structurally unconscious, or merely preconscious due to IIB? If a person says they became aware of the process after attending to it, how could we objectively verify their claim?

Here, I lay out a schematic for how IIB could be tested rigorously (Figure 2). The thrust is to induce people to better attend to their internal events (through process-specific cues or general training), and then test whether they can more accurately report on a target process in a way that most plausibly came from increased conscious awareness of the process.

**How to test for IIB: roadmap and example**

| *Roadmap* | *Example* |
|---|---|
| Step 1: Induce attention to the (component of) the target process (via process-specific cue or general training). | Participants undergo either a weeklong training on awareness of internal experience (e.g. through regular guided interviews), or a control training. |
| Step 2: Elicit self-reports about the process, and compare to objective evidence. | Participants complete a complex multi-attribute choice task, and are subsequently asked to report how much weight they placed on each attribute. Do people who undergo the training report weights that correlate more strongly with the weights extracted via statistical modeling? |
| Step 3: Provide evidence against improved inference as an explanation. | Other "observer" participants undergo the training, and then, after viewing *another* person's choices, judge how much weight *that* person placed on each attribute. Show that the training doesn't improve observers' ability to infer others' attribute weights. |

*Figure 2*. Conceptual roadmap of how to test whether a (component of) a mental process is unconscious due to internal inattentional blindness, and can be brought into awareness via attention. Any test must (1) induce people to attend to the target component of the mental process (by either cueing people to the process directly or providing general attentional training); (2) elicit people's self-reports about the target process, and compare their reports to objective evidence; and (3) if the attention induction improves the match between people's self-reports and the objective evidence, provide evidence that the improvement did not come from improved inference about the process (as posited by, e.g., Bem (1967); Carruthers (2009); Nisbett and Wilson (1977)).

To take one of myriad potential examples, consider the "mere exposure effect": the finding that repeated exposure to a novel stimulus makes people like it more (Zajonc, 1968). This effect is extremely robust, and yet people are largely unaware of it, typically denying an influence of familiarity and instead attributing their positive

attitude to some irrelevant property of the familiar stimulus (Bornstein & Craver-Lemley, 2016). To test whether the internal events underlying this effect (e.g. the activation of a sense of "familiarity" while making liking judgments) are unconscious due to IIB, we would induce people to attend to those events and then test whether they can more accurately report the influence of familiarity on their liking judgments.

This simple description, of course, belies complex methodological issues. For one, we need to know how to induce people to internally attend to a target process – a feat which, according to practitioner reports, can be very difficult. We also need to know how to objectively measure the accuracy of people's self-reports about the process (Varela, 1996).

Most challengingly, we need to show that any improvements in reporting accuracy are due to increased conscious access to the mental process, as opposed to an improved theory or inference about the process (Nisbett & Wilson, 1977). This issue is particularly delicate. We know that people make inferences and build complex theories about the underlying causes of people's behavior, using observation, reasoning, and cultural knowledge (Kelley, 1967; Tenenbaum, Griffiths, & Kemp, 2006); and we know that they apply this cognitive machinery towards themselves, inferring their own mental processes through self-observation (Bem, 1972; Carruthers, 2009; Wilson, 2004). Moreover, classic evidence against introspection revolves around this confound, showing that many apparent instances of accurate introspection are actually instances either of inference from self-observation or of "incidentally correct employment of *a priori* causal theories" (Nisbett & Wilson, 1977). The same problem applies here: In order to show that internal attention improves conscious awareness of a mental process, we need to show that increases in the accuracy of people's self-reports is not due to an improved inference or theory about the process. This is perhaps the largest obstacle when developing experimental tests of IIB.[9]

---

[9] Note that, here, we are in an opposite position from most external inattentional blindness experiments. In most IB experiments, the target stimulus is by default conscious (e.g. a shape flashed for 200ms, or a gorilla walking across a screen); the test condition renders it unconscious by diverting attention; and delicate experimental work is required to show that, in the test condition, people are failing to report it because of inattention (as opposed to some other reason). Here, our targets are by

I will discuss each issue in turn, with the goal, not to develop a specific experiment or overcome all possible methodological obstacles, but rather to illustrate that IIB is testable and provide a roadmap for how to do it.

### 6.1 Inducing people to attend to a mental process

The first challenge facing empirical tests of IIB is inducing people to attend to a target mental process. There are two basic approaches: directly cue people to attend to the target process/internal event, or train people to attend to internal events in a process-general way.

Past studies have employed both approaches. The studies by Hahn et al. on implicit attitude awareness, for instance, take the former approach; they describe what implicit attitudes are, direct people to attend to them, and measure people's subsequent ability to report them accurately (Hahn & Gawronski, 2019; Hahn et al., 2014). The studies with guided coaching, in contrast, take the latter approach; they train people to attend broadly to their internal events (avoiding distractions, theorizing, and other common pitfalls of introspection), and then measure people's subsequent accuracy at reporting things like internal speech (Hurlburt et al., 2016; Kühn et al., 2014).

In practice, most applied modalities combine the two approaches. For example, in therapies like CBT and DBT, patients practice attending to internal events in a domain-general way (e.g. through thought journaling), and also are cued to attend to specific thought patterns or emotional processes (Beck, 1979; Linehan, 2018; Segal et al., 2018). Similarly, meditators practice focusing attention in general, and are also cued to attend to specific sensations (like mental images or inner speech; Goldstein (2017); Kabat-Zinn (1994); Young (2016)).

What kinds of process-general training would be the best candidates for bringing unconscious mental processes into awareness? There are a variety of modalities

---

default unconscious (i.e. unconscious higher mental processes); our desired test condition would render them conscious by focusing attention; and delicate experimental work is required to show that, in the test condition, people are succeeding at reporting them because of improved attention. An additional prediction of the IIB hypothesis is that conscious mental processes can be rendered unconscious by diverting internal attention. However, since my focus is on whether unconscious processes can become conscious, I will not develop this idea further here.

involving internal attention training, and reviewing them all is outside the scope of this paper. Two particularly promising approaches would be meditation methods focused on non-judgmentally attending to internal cognitions (like noticing thoughts, mental images, inner speech, etc.; Young (2016)), and guided interview methods (Hurlburt et al., 2016; Hurlburt & Heavey, 2001; Petitmengin, 2006). There are also promising methods which train people by repeatedly giving them direct, immediate feedback on the accuracy of their internal attention and letting them learn implicitly how to control it (Bernstein & Zvielli, 2014). Which, if any, of these training methods is successful at reducing IIB is a question for future empirical work.

Motivated by the difficulties of directing internal attention, though, we can identify three features that any internal attention training should have to be successful. It should give people practice focusing their internal attention on increasingly subtle experiences while simultaneously attending to external tasks. It should teach them to notice when they are *not* attending to direct experience (when they are theorizing, making inferences, rationalizing, etc.), and to refocus. And it should help them overcome the motivation to turn away from unpleasant internal discoveries, perhaps by helping them observe themselves with equanimity or compassion (Kabat-Zinn, 2015; Young, 2016).

### 6.2 Measuring the accuracy of people's self-reports about a process

The second challenge facing empirical tests of IIB is to determine whether the attention induction succeeded – i.e. whether people subsequently become more accurate at reporting their internal events or processes. This requires comparing people's reports to objective, external evidence for the existence of an internal event/process, and testing whether people who receive an attention induction give reports that, on average, better match the objective evidence (Varela, 1996).

Ways to do this abound. For instance, in a mere exposure effect paradigm, we have objective, experimental evidence that people's liking judgments are influenced by familiarity with the stimuli (Bornstein & Craver-Lemley, 2016). If people who undergo internal attention training then more often report that, in a mere exposure effect

experiment, their liking judgments are being influenced by familiarity, this would be evidence that the role of familiarity was only incidentally unconscious and can be made conscious via internal attention. Or another example: In multi-attribute choice paradigms (e.g. choosing between cars which vary simultaneously on price, looks, horsepower, gas mileage, etc), we can determine from people's choices the weights they're placing on each attribute (Bhatia & Stewart, 2018). If people who undergo internal attention training report weights that correlate better with the observed weights, that would be evidence that the decision process can be made more conscious via internal attention.

These examples are two of many. Any mental process posited by psychologists has presumably been posited because of objective evidence; it was in part the mismatch between people's reports and the objective evidence that led researchers to distrust introspection in the first place (Nisbett & Wilson, 1977). If, after training, people's self-reports better match the objective evidence for a target process, this would suggest that the process was only incidentally unconscious and could be brought into awareness.

There is an important nuance here. It is easier to experimentally establish the existence of a mental process at a group level – i.e. to show that people, on average, like familiar stimuli more than unfamiliar ones, or use "ease of generating examples" as a factor in frequency judgment. It is more difficult to prove what happened in any one person's head. The existence of many processes can only be robustly observed in between-subjects manipulations; and even for effects that can be observed with within-subject manipulations (e.g. a participant in a mere exposure effect paradigm who reports liking the common stimuli more than the uncommon ones), it is hard to determine whether the target mental process definitely occurred in that individual (maybe that participant just happened to prefer those stimuli over the others). Of course, there are contexts where we can get good objective evidence about what's going on in an individual's head. For instance, we can identify the weights a person places on attributes in multi-attribute choice (Bhatia & Stewart, 2018); measure a person's implicit attitudes (Hahn et al., 2014); or use process-tracing methods (like eye-tracking)

to gain more fine-grained information about the person's process (Schulte-Mecklenbeck, Kühberger, & Johnson, 2011). But establishing that a mental process happened in an individual will inevitably be noisier and harder than establishing that a mental process happens on average.

Fortunately, both types of evidence – individual- and group-level – can be used to test for IIB. In contexts where we can get individual-level evidence for the existence of an internal process/event, we can test whether internal attention improves the match between each individual's observed, idiosyncratic process and their self-report of that process (e.g. testing whether training improves the correlation between self-reported and observed attribute weights in a multi-attribute choice task). And in contexts where we only know that a process happens on average (but can't determine whether it happens in any one person), we can test whether internal attention improves people's ability, on average, to accurately report the process we know to be occurring on average (e.g. testing whether training improves the percentage of people who report using "ease of generating examples" to make frequency judgments, a process which we know to be occurring on average). These approaches can be complementary, with individual-level tests being more precise but group-level tests enabling an examination of more diverse processes.

### 6.3 Showing that improvements in introspective accuracy are due to increased conscious access, rather than improved inference or theorizing

A third challenge facing any experimental test of IIB is to show that accurate reporting of the target process is due to increased conscious access to the process, rather than improved inference or theorizing about it. As described above, people have lay theories about how their minds work, and can make inferences about their mental processes without having conscious access to those processes (Bem, 1967; Carruthers, 2009; Gopnik, 1993; Nisbett & Wilson, 1977; Wilson, 2004). Nisbett and Wilson (1977) rightfully argued that, to claim that accurate reports are due to introspection, a central challenge is to rule out that the reports came from these kind of self-observing

inferential processes. The same is true here: To claim that *improvements* in accurate reporting came from increased conscious awareness, we would need to show that the improvements did not come from better self-observation / inference.

Though an inference account would be difficult to completely rule out, there are several ways to provide evidence against it. One approach would be to use attention inductions whose content is completely unrelated to the target process. It is not obvious how a course on mindfulness meditation, for example, would provide new observations sufficient to infer that familiarity breeds liking in a mere exposure paradigm, or to infer the weight placed on "gas mileage" in a multi-attribute choice task. So if these interventions successfully improved introspective accuracy, improved inference would be an unlikely explanation.

Another approach would be to test for awareness of aspects of the target process that would be difficult to infer without conscious access. For instance, the mere exposure effect does not occur when the stimuli presentations are clumped together homogenously (i.e. when all presentations of a stimulus occur consecutively; Bornstein and Craver-Lemley (2016)). This boundary condition would be extremely difficult to infer from lay theorizing; yet, a person who attained conscious access to the process underlying their liking judgments would be able to correctly report that familiarity did not, in that context, play a role.

Finally, researchers could adopt the suggestion of Nisbett and Wilson (1977): Compare people's self-reports to the inferences of observers. Nisbett & Wilson recommended giving "observer" participants a vague description of the task and seeing if they can infer people's mental processes as well as the people themselves. Researchers could go a step further by matching each observer with an actual participant, and giving the observer a description of as many observable facts about the participant as possible: their full experience of the task, their demographic information, even a video of them performing the task. Participants and observers could both receive the attention induction. If the induction improves people's ability to report their own process but does not improve observers' ability to infer another person's process, that

would provide further evidence that internally attending to the process improves accuracy by bringing it into conscious awareness.

Of course, even using the most stringent controls, it would be difficult to rule out that the training improved people's motivation or ability to infer their own mental processes through some internal self-observation technique that is different from gaining conscious access to the process itself. Though these fine-grained distinctions matter theoretically, at a practical level they start to become less important. If internal attentional training substantially improves people's ability to accurately report on their own unconscious higher mental processes in a way that observers cannot, this finding would suggest that people can learn to access a "fount of privileged knowledge" (Bem, 1967) about their own mental processes. Moreover, this finding would still significantly advance our understanding of the potential of the conscious mind, and provide a foundation for improving applied modalities like therapy and mindfulness.

## 7 Implications of the IIB hypothesis

If applied practitioners are right, and important high-level mental processes can be demonstrably brought into awareness via attention, this finding would have significant ramifications for basic experimental psychology.

For one, it would start to answer a fundamental question about the mind: how much of our own mental lives we can directly perceive. This question shows up in a diverse array of influential theories. As discussed above, it shows up directly in the work of Nisbett & Wilson, who argued that people have little introspective access to their mental processes "no matter how hard we try" (Nisbett & Wilson, 1977; Wilson, 2004). But similar sentiments show up indirectly in other accounts. In the introduction to his bestseller *Thinking Fast & Slow*, Kahneman (2011) writes:

> When you are asked what you are thinking about, you can normally answer. You believe you know what goes on in your mind, which often consists of one conscious thought leading in an orderly way to another. But that is not the only way the mind works, nor indeed is that the typical way. Most impressions and thoughts arise in your conscious experience without your

> knowing how they got there... The mental work that produces impressions, intuitions, and many decisions goes on in silence in our mind.

A similar dual-process account is given by Haidt (2012) in his bestseller *The Righteous Mind*:

> The rider is our conscious reasoning – the stream of words and images of which we are fully aware. The elephant is the other 99 percent of mental processes – the ones that occur outside of awareness but that actually govern most of our behavior.

Dual-process accounts, like those proposed by Kahneman, Haidt, and others, do not depend on "System 1" being permanently unconscious (Evans, 2003). Nonetheless, these introductions would have to be rewritten if it turned out that people, with training, could become directly conscious of many of the hidden processes producing their intuitions, judgments, and choices. The same applies to theories of implicit attitudes (Greenwald & Banaji, 1995; Hahn et al., 2014), automatic social cognition (Bargh, 2013; Bargh & Morsella, 2008; Bargh & Williams, 2006), moral psychology (Greene, 2013; Haidt, 2001), and more (Hassin et al., 2004).

The IIB hypothesis, if it were true, would also have implications for accounts of "cognitive illusions" (Pohl, 2016). For instance, consider the "illusion of conscious will" (Wegner, 2017). Much evidence suggests that people's sense of consciously willing an action is dissociable from the actual mental events producing the action (Libet, 1985; Wegner, 2003). Building on these findings, Wegner (2003, 2017) argues that people do not access veridical representations of intention or action initiation, and instead infer conscious causation post-hoc – i.e. they experience an illusion of conscious will. Though Wegner's account does not depend on the illusion being permanent, the interpretation of these data would be quite different if people stopped showing the illusion after training. The same is true for theories of choice blindness (Johansson et al., 2005; Petitmengin et al., 2013), constructed preferences (Slovic, 1995), and other cognitive illusions (Chater, 2018; Pohl, 2016).

Despite the potentially-widespread theoretical impact of the IIB hypothesis, it has received almost no direct investigation. Nearly all the evidence for introspective-inaccessibility in high-level cognition boils down to the fact that untrained participants cannot accurately report their mental processes (Hurlburt & Heavey, 2001; Nisbett & Wilson, 1977; Petitmengin, 2006). This evidence has been taken as dispositive because the consciousness or unconsciousness of a process is, by default, assumed to be a structural feature of the mind – i.e. if lay people can't access a process, it's inaccessible (Bargh & Morsella, 2008).[10] But if practitioner reports are to be taken seriously, this evidence is highly inconclusive; it would be akin to concluding that mathematical truths are inherently inaccessible because eight-year-olds cannot comprehend calculus. Another way to put it is that this accessible-or-not dichotomy is not true of external perception; external perception is highly dependent on attention. There are some things that are permanently unconscious and others that are easy to see, but many stimuli are in the middle – you only become conscious of them if you look carefully. The same is plausibly true for internal perception – and doubly so, since the internal cues are inherently weaker and internal attention is plausibly more difficult to control. Accounts which invoke the scope of consciousness in the mind cannot afford to ignore the possibility of IIB, and of a labile, skill-based boundary between conscious and unconscious.[11]

On the flip side, applied modalities like cognitive therapies and mindfulness trainings cannot afford to go on assuming the existence of IIB without testing it. These

---

[10] Researchers sometimes accede the possibility of increasing the scope of consciousness through training, but it is rarely given serious weight. For instance, Nisbett and Wilson (1977) acknowledge that their evidence does not "suffice to show that people could never be accurate" when reporting their processes, and that "interrupting a process at the very moment it was occurring, alerting subjects to pay careful attention to their cognitive processes, coaching them in introspective procedures, and so on" could potentially improve introspective access. But they quickly dismiss such possibilities as "ecologically meaningless".

[11] Another set of psychological accounts that would be impacted by the existence of genuine introspection training are ones that rely largely on self-report: for instance, positive psychology (Peterson, 2006), personality psychology (Paulhus, Vazire, et al., 2007), symptom-focused clinical psychology, and studies of explicit attitudes (Ajzen, 1991). These disciplines often rely on people's ability to report features of their mental lives – their happiness, dispositions, attitudes, internal experiences, and so on. If people can become provably better at noticing and reporting internal events, they could plausibly learn to answer these questions more accurately.

modalities are enormously popular; 50% of American households have someone visit a therapist each year (Chamberlin, 2004), over 40% of Americans report meditating at least once a week (Masci & Hackett, 2017), and there is a burgeoning industry of other mental training practices aimed, in part, at improving "experiential access" to mental processes (Dahl et al., 2020). Though there is much evidence for the overall efficacy of these techniques, their mechanistic claims sometimes far outpace their evidence base; they report bringing unconscious processes into awareness, but almost never test those claims objectively. If the IIB hypothesis is right, it could offer a cognitively-grounded framework for developing, improving, and validating introspection trainings. And if it is wrong, then a key aspect of these modalities must be reconsidered.

Finally, as discussed above, if the IIB hypothesis is right, it could impact the practice of psychological science. It is a poorly-kept secret that much of our science is guided by scientists' introspection. To quote Jack et al. (2003):

> An informal reliance on introspective evidence is ubiquitous in psychology and cognitive science. It generates many of the hypotheses that psychologists seek to test using objective sources of evidence, it underlies their understanding of cognitive tasks or 'task analysis', and it frequently informs the questions and objections they offer as referees. Introspective understanding even forms the basis of many of the categories used to describe branches of psychological research (e.g. 'attention', 'episodic memory', 'awareness').

Even though our science does not rest on introspective evidence, it is often informally guided by it (Locke, 2009; Schooler, 2002a). Just as turning on more lights would dramatically accelerate the pace of finding your lost keys, expanding conscious access to mental processes could significantly speed up the pace of psychological discovery.

## 8 Conclusion

Just as people can miss perceptual events due to external inattention, so may they be blind to internal events – like those constituting high-level mental processes – due to

internal inattention. The existence of internal inattentional blindness, and the possibility of overcoming it through training, are widely assumed in successful applied psychological modalities and widely reported by practitioners; yet these possibilities have rarely been explored experimentally, or taken seriously by basic theorists. Rigorously demonstrating the existence of IIB could open a new chapter both in the development of psychological interventions, and in our scientific understanding of the scope of conscious awareness.

# 9 References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179–211.

Amir, I., Ruimi, L., & Bernstein, A. (2021). Simulating thoughts to measure and study internal attention in mental health. *Scientific reports*, *11*(1), 1–17.

Awh, E., Vogel, E. K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208.

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences*, *6*(1), 47–52.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206.

Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., ... Williams, J. M. G. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment*, *15*(3), 329–342.

Bargh, J. A. (2013). *Social psychology and the unconscious: The automaticity of higher mental processes*. Psychology Press.

Bargh, J. A., & Morsella, E. (2008). The unconscious mind. *Perspectives on psychological science*, *3*(1), 73–79.

Bargh, J. A., & Williams, E. L. (2006). The automaticity of social life. *Current directions in psychological science*, *15*(1), 1–4.

Beck, A. T. (1979). *Cognitive therapy of depression*. Guilford press.

Beck, A. T. (1991). Cognitive therapy: A 30-year retrospective. *American psychologist*, *46*(4), 368.

Beck, A. T., & Haigh, E. A. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual review of clinical psychology*, *10*, 1–24.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological review*, *74*(3), 183.

Bem, D. J. (1972). Self-perception theory. In *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Elsevier.

Bernstein, A., & Zvielli, A. (2014). Attention feedback awareness and control training (a-fact): Experimental test of a novel intervention paradigm targeting attentional bias. *Behaviour research and therapy*, *55*, 18–26.

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., ... others (2004). Mindfulness: a proposed operational definition. *Clinical psychology: Science and practice*, *11*(3), 230.

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

Block, N. (2005). Two neural correlates of consciousness. *Trends in cognitive sciences*, *9*(2), 46–52.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in cognitive sciences*, *15*(12), 567–575.

Bornstein, R. F., & Craver-Lemley, C. (2016). Mere exposure effect. *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory*, *256*.

Bredemeier, K., & Simons, D. J. (2012). Working memory and inattentional blindness. *Psychonomic Bulletin & Review*, *19*(2), 239–244.

Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We see more than we can report: "cost free" color phenomenality outside focal attention. *Psychological science*, *25*(7), 1394–1403.

Brown, D., Forte, M., & Dysart, M. (1984). Visual sensitivity and mindfulness meditation. *Perceptual and motor skills*, *58*(3), 775–784.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, *84*(4), 822.

Caporale-Berkowitz, N. A., Boyer, B. P., Lyddy, C. J., Good, D. J., Rochlen, A. B., &

Parent, M. C. (2021). Search inside yourself: investigating the effects of a widely adopted mindfulness-at-work development program. *International Journal of Workplace Health Management*.

Carlson, E. N. (2013). Overcoming the barriers to self-knowledge: Mindfulness as a path to seeing yourself as you really are. *Perspectives on Psychological Science*, *8*(2), 173–186.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and brain sciences*, *32*(2), 121–138.

Castonguay, L. G., & Hill, C. E. (2007). Insight in psychotherapy.

Chamberlin, J. (2004). Survey says: More americans are seeking mental health treatment. *Monitor on Psychology*, *35*(7).

Chan, D., & Woollacott, M. (2007). Effects of level of meditation experience on attentional focus: is the efficiency of executive or orientation networks improved? *The Journal of Alternative and Complementary Medicine*, *13*(6), 651–658.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80–94.

Chater, N. (2018). *Mind is flat: The remarkable shallowness of the improvising brain*. Yale University Press.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, *62*, 73–101.

Chun, M. M., & Johnson, M. K. (2011). Memory: Enduring traces of perceptual and reflective attention. *Neuron*, *72*(4), 520–535.

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in cognitive sciences*, *16*(8), 411–417.

Dahl, C. J., Wilson-Mendenhall, C. D., & Davidson, R. J. (2020). The plasticity of well-being: A training-based framework for the cultivation of human flourishing. *Proceedings of the National Academy of Sciences*, *117*(51), 32197–32206.

De Brigard, F. (2012). The role of attention in conscious recollection. *Frontiers in Psychology*, *3*, 29.

De Brigard, F., & Prinz, J. (2010). Attention and consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 51–59.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications. *Characterizing consciousness: From cognition to the clinic?*, 55–84.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences*, *10*(5), 204–211.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

DeRubeis, R. J., Webb, C. A., Tang, T. Z., & Beck, A. T. (2010). Cognitive therapy. In *Handbook of cognitive-behavioral therapies, 3rd ed* (pp. 277–316). New York, NY, US: Guilford Press.

Dixon, M. L., Fox, K. C., & Christoff, K. (2014). A framework for understanding the relationship between externally and internally directed cognition. *Neuropsychologia*, *62*, 321–330.

Drew, T., Võ, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological science*, *24*(9), 1848–1853.

Dunne, J. D., Thompson, E., & Schooler, J. (2019). Mindful meta-awareness: Sustained

and non-propositional. *Current Opinion in Psychology*, *28*, 307–311.

Dux, P. E., & Marois, R. (2009). The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics*, *71*(8), 1683–1700.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* the MIT Press.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, *7*(10), 454–459.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition*, *129*(2), 292–308.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). *Metacognition: computation, biology and function.* The Royal Society.

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

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Fortney, M. (2020). Directing internal attention towards ongoing thought. *Consciousness and Cognition*, *85*, 103025.

Fox, K. C., Zakarauskas, P., Dixon, M., Ellamil, M., Thompson, E., & Christoff, K. (2012). Meditation experience predicts introspective accuracy. *PloS one*, *7*(9), e45370.

Frank, J. D., & Frank, J. B. (1993). *Persuasion and healing: A comparative study of psychotherapy.* JHU Press.

Froese, T., Gould, C., & Seth, A. (2011). Validating and calibrating first-and second-person methods in the science of consciousness. *Journal of Consciousness*

*Studies*, *18*(2), 38.

Gehring, W. J., Bryck, R. L., Jonides, J., Albin, R. L., & Badre, D. (2003). The mind's eye, looking inward? in search of executive control in internal attention shifting. *Psychophysiology*, *40*(4), 572–585.

Ghasemipour, Y., Robinson, J. A., & Ghorbani, N. (2013). Mindfulness and integrative self-knowledge: Relationships with health-related variables. *International Journal of Psychology*, *48*(6), 1030–1037.

Gibbons, F. X. (1983). Self-attention and self-report: The "veridicality" hypothesis. *Journal of Personality*, *51*(3), 517–542.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Goldstein, J. (2017). *The experience of insight: A simple and direct guide to buddhist meditation*. Shambhala Publications.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain sciences*, *16*(1), 1–14.

Grant, A. M., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality: an international journal*, *30*(8), 821–835.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of cognitive neuroscience*, *15*(8), 1176–1194.

Grosse Holtforth, M., Castonguay, L. G., Boswell, J. F., Wilson, L. A., Kakouros, A. A., & Borkovec, T. D. (2007). Insight in cognitive-behavioral therapy.

Hadash, Y., & Bernstein, A. (2019). Behavioral assessment of mindfulness: defining features, organizing framework, and review of emerging methods. *Current opinion in psychology*, *28*, 229–237.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hanley, A. W., & Garland, E. L. (2017). Clarity of mind: Structural equation modeling of associations between dispositional mindfulness, self-concept clarity and psychological well-being. *Personality and individual differences*, *106*, 334–339.

Hassin, R. R., Uleman, J. S., & Bargh, J. A. (2004). *The new unconscious*. Oxford university press.

Hayes, S. C., & Hofmann, S. G. (2017). The third wave of cognitive behavioral therapy and the rise of process-based care. *World Psychiatry*, *16*(3), 245.

Hill, C. E., & Knox, S. (2008). Facilitating insight in counseling and psychotherapy. In *Handbook of counseling psychology, 4th ed.* (pp. 284–302). Hoboken, NJ, US: John Wiley & Sons, Inc.

Hirst, W., Spelke, E. S., Reaves, C. C., Caharack, G., & Neisser, U. (1980). Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General*, *109*(1), 98.

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, *36*(5), 427–440.

Hurlburt, R., & Schwitzgebel, E. (2011). *Describing inner experience?: Proponent meets skeptic*. Mit Press.

Hurlburt, R. T., & Akhter, S. A. (2006). The descriptive experience sampling method. *Phenomenology and the Cognitive Sciences*, *5*(3-4), 271–301.

Hurlburt, R. T., Alderson-Day, B., Kühn, S., & Fernyhough, C. (2016). Exploring the

ecological validity of thinking on demand: neural correlates of elicited vs. spontaneously occurring inner speech. *PLoS one*, *11*(2), e0147932.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: describing inner experience. *Trends in cognitive sciences*, *5*(9), 400–403.

Hyman, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, *24*(5), 597–607. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.1638)

Jack, A., et al. (2003). Why trust the subject? *Journal of consciousness studies*, *10*(9-10), v–xx.

Jha, A. P., Krompinger, J., & Baime, M. J. (2007). Mindfulness training modifies subsystems of attention. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 109–119.

Jo, H.-G., Hinterberger, T., Wittmann, M., & Schmidt, S. (2015). Do meditators have higher awareness of their intentions to act? *Cortex*, *65*, 149–158.

Jo, H.-G., Wittmann, M., Borghardt, T. L., Hinterberger, T., & Schmidt, S. (2014). First-person approaches in neuroscience of consciousness: brain dynamics correlate with the intention to act. *Consciousness and Cognition*, *26*, 105–116.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Kabat-Zinn, J. (1994). *Wherever you go, there you are: Mindfulness meditation in everyday life*. Hachette Books.

Kabat-Zinn, J. (2015). Mindfulness. *Mindfulness*, *6*(6), 1481–1483.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kegan, R., Kegan, L. L. L. R., & Lahey, L. L. (2009). *Immunity to change: How to overcome it and unlock potential in yourself and your organization*. Harvard Business Press.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium*

*on motivation.*

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic bulletin & review*, *20*(2), 228–242.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, *11*(1), 16–22.

Koole, S. L., Govorun, O., Cheng, C. M., & Gallucci, M. (2009). Pulling yourself together: Meditation promotes congruence between implicit and explicit self-esteem. *Journal of Experimental Social Psychology*, *45*(6), 1220–1226.

Kühn, S., Fernyhough, C., Alderson-Day, B., & Hurlburt, R. T. (2014). Inner experience in the scanner: can high fidelity apprehensions of inner experience be integrated with fmri? *Frontiers in Psychology*, *5*, 1393.

Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends in cognitive sciences*, *7*(1), 12–18.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, *15*(8), 365–373.

Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, *4*(4), 501–506.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of cognitive neuroscience*, *24*(1), 61–79.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and brain sciences*, *8*(4), 529–539.

Linehan, M. M. (2018). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Publications.

Locke, E. A. (2009). It's time we brought introspection out of the closet. *Perspectives on Psychological Science*, *4*(1), 24–25.

Logan, G. D., Cox, G. E., Annis, J., & Lindsey, D. R. (2021). The episodic flanker effect: Memory retrieval as attention turned inward. *Psychological Review*,

*128*(3), 397.

Lückmann, H. C., Jacobs, H. I., & Sack, A. T. (2014). The cross-functional role of frontoparietal regions in cognition: internal attention as the overarching mechanism. *Progress in neurobiology*, *116*, 66–86.

Lush, P., Naish, P., & Dienes, Z. (2016). Metacognition of intentions in mindfulness and hypnosis. *Neuroscience of Consciousness*, *2016*(1).

Lutz, A., Jha, A. P., Dunne, J. D., & Saron, C. D. (2015). Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *American Psychologist*, *70*(7), 632.

Lutz, A., Lachaux, J.-P., Martinerie, J., & Varela, F. J. (2002). Guiding the study of brain dynamics by using first-person data: synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proceedings of the national academy of sciences*, *99*(3), 1586–1591.

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in cognitive sciences*, *12*(4), 163–169.

Macdonald, J. S., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics*, *73*(6), 1780–1789.

Mack, A., Rock, I., et al. (1998). *Inattentional blindness*. MIT press.

MacLean, K. A., Ferrer, E., Aichele, S. R., Bridwell, D. A., Zanesco, A. P., Jacobs, T. L., ... others (2010). Intensive meditation training improves perceptual discrimination and sustained attention. *Psychological science*, *21*(6), 829–839.

Maoz, U., Mudrik, L., Rivlin, R., Ross, I., Mamelak, A., & Yaffe, G. (2015). On reporting the onset of the intention to move. *Surrounding free will: philosophy, psychology, neuroscience (Mele A, ed)*, 184–202.

Masci, D., & Hackett, C. (2017). Meditation is common across many religious groups in the u.s.

McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. Guilford Press.

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional

blindness. *Consciousness and cognition*, *15*(3), 620–627.

Morales, J. (2021, May). Introspection Is Signal Detection. *The British Journal for the Philosophy of Science*. (Publisher: The University of Chicago Press)

Most, S. B., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What you see is what you set: sustained inattentional blindness and the capture of awareness. *Psychological review*, *112*(1), 217.

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: beyond sustained internal attention. *Trends in cognitive sciences*, *21*(6), 449–461.

Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159–168.

Nakajima, M., Takano, K., & Tanno, Y. (2017). Adaptive functions of self-focused attention: Insight and depressive and anxiety symptoms. *Psychiatry research*, *249*, 275–280.

Nakajima, M., Takano, K., & Tanno, Y. (2019). Mindfulness relates to decreased depressive symptoms via enhancement of self-insight. *Mindfulness*, *10*(5), 894–902.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive psychology*, *7*(4), 480–494.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Nyklíček, I., & Denollet, J. (2009). Development and evaluation of the balanced index of psychological mindedness (bipm). *Psychological Assessment*, *21*(1), 32.

Nyklíček, I., Zonneveld, R., & Denollet, J. (2020). Introspective interest and insight in the context of mindfulness-based stress reduction: a randomized trial. *Mindfulness*, *11*(9), 2176–2188.

Paulhus, D. L., Vazire, S., et al. (2007). The self-report method. *Handbook of research methods in personality psychology*, *1*(2007), 224–239.

Pearson, J., Rademaker, R. L., & Tong, F. (2011). Evaluating the mind's eye: the

metacognition of visual imagery. *Psychological Science*, *22*(12), 1535–1542.

Peterson, C. (2006). *A primer in positive psychology*. Oxford university press.

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive sciences*, *5*(3), 229–269.

Petitmengin, C. (2009). *Ten Years of Viewing From Within: The Legacy of Francisco Varela*. Imprint Academic.

Petitmengin, C., Baulac, M., & Navarro, V. (2006). Seizure anticipation: are neurophenomenological approaches able to detect preictal symptoms? *Epilepsy & Behavior*, *9*(2), 298–306.

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in nisbett and wilson's findings? a first-person access to our cognitive processes. *Consciousness and cognition*, *22*(2), 654–669.

Pinker, S. (1997). *How the mind works*. Princeton University Press.

Pohl, R. F. (2016). *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory*. Psychology Press.

Rademaker, R. L., & Pearson, J. (2012). Training visual imagery: Improvements of metacognition, but not imagery strength. *Frontiers in psychology*, *3*, 224.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an rsvp task: An attentional blink? *Journal of experimental psychology: Human perception and performance*, *18*(3), 849.

Remmers, C., Zimmermann, J., Buxton, A., Unger, H.-P., Koole, S. L., Knaevelsrud, C., & Michalak, J. (2018). Emotionally aligned: Preliminary results on the effects of a mindfulness-based intervention for depression on congruence between implicit and explicit mood. *Clinical psychology & psychotherapy*, *25*(6), 818–826.

Rensink, R. A. (2002). Change detection. *Annual review of psychology*, *53*(1), 245–277.

Richards, A., Hannon, E. M., & Derakshan, N. (2010). Predicting and manipulating the incidence of inattentional blindness. *Psychological research*, *74*(6), 513–523.

Schooler, J. W. (2002a). Establishing a legitimate relationship with introspection:

Response to jack and roepstorff. *Trends in cognitive sciences*, *6*(9), 371–372.

Schooler, J. W. (2002b). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in cognitive sciences*, *6*(8), 339–344.

Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (2011). A handbook of process tracing methods for decision research: A critical review and user's guide.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, *61*(2), 195.

Schwarz, N., & Vaughn, L. A. (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information.

Schwitzgebel, E. (2004). Introspective training apprehensively defended: Reflections on titchener's lab manual. *Journal of Consciousness Studies*, *11*(7-8), 58–76.

Seegmiller, J. K., Watson, J. M., & Strayer, D. L. (2011). Individual differences in susceptibility to inattentional blindness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 785.

Segal, Z. V., Williams, M., & Teasdale, J. (2018). *Mindfulness-based cognitive therapy for depression*. Guilford Publications.

Shapiro, K. L., Raymond, J., & Arnell, K. (1997). The attentional blink. *Trends in cognitive sciences*, *1*(8), 291–296.

Shea, N., & Frith, C. D. (2019). The global workspace needs metacognition. *Trends in cognitive sciences*, *23*(7), 560–571.

Shear, J., & Jevning, R. (1999). Pure consciousness: Scientific exploration of meditation techniques. *Journal of consciousness studies*, *6*(2-3), 189–210.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, *28*(9), 1059–1074.

Simons, D. J., & Jensen, M. S. (2009). The effects of individual differences and task difficulty on inattentional blindness. *Psychonomic Bulletin & Review*, *16*(2), 398–403.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a

real-world interaction. *Psychonomic Bulletin & Review*, *5*(4), 644–649.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences*, *9*(1), 16–20.

Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., & Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biol*, *5*(6), e138.

Slovic, P. (1995). The construction of preference. *American psychologist*, *50*(5), 364.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to nisbett and wilson.

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860.

Spelke, E., Hirst, W., & Neisser, U. (1976). Skills of divided attention. *Cognition*, *4*(3), 215–230.

Stein, D., & Grant, A. M. (2014). Disentangling the relationships among self-reflection, insight, and subjective well-being: The role of dysfunctional attitudes and core self-evaluations. *The Journal of psychology*, *148*(5), 505–522.

Strick, M., & Papies, E. K. (2017). A brief mindfulness exercise promotes the correspondence between the implicit affiliation motive and goal setting. *Personality and Social Psychology Bulletin*, *43*(5), 623–637.

Tang, Y.-Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., ... others (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences*, *104*(43), 17152–17156.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

(Thera), N. (1972). *The power of mindfulness*. Unity Press.

Timulak, L., & McElvaney, R. (2013). Qualitative meta-analysis of insight events in psychotherapy. *Counselling Psychology Quarterly*, *26*(2), 131–150.

Treves, I. N., Tello, L. Y., Davidson, R. J., & Goldberg, S. B. (2019). The relationship

between mindfulness and objective measures of body awareness: A meta-analysis. *Scientific reports*, *9*(1), 1–12.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, *79*(4), 281.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

Vago, D. R., & David, S. A. (2012). Self-awareness, self-regulation, and self-transcendence (s-art): a framework for understanding the neurobiological mechanisms of mindfulness. *Frontiers in human neuroscience*, *6*, 296.

Van Boxtel, J. J., Tsuchiya, N., & Koch, C. (2010). Consciousness and attention: on sufficiency and necessity. *Frontiers in Psychology*, *1*, 217.

Van Ede, F., Board, A. G., & Nobre, A. C. (2020). Goal-directed and stimulus-driven selection of internal representations. *Proceedings of the National Academy of Sciences*, *117*(39), 24590–24598.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of consciousness studies*, *3*(4), 330–349.

Wampold, B. E., Imel, Z. E., Bhati, K. S., & Johnson-Jennings, M. D. (2007). Insight as a Common Factor. In *Insight in psychotherapy.* (pp. 119–139). Washington, DC, US: American Psychological Association.

Ward, E. J., & Scholl, B. J. (2015). Inattentional blindness reflects limitations on perception, not memory: Evidence from repeated failures of awareness. *Psychonomic Bulletin & Review*, *22*(3), 722–727.

Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends in cognitive sciences*, *7*(2), 65–69.

Wegner, D. M. (2017). *The illusion of conscious will*. MIT press.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of nisbett and wilson and of bem.

Wilson, T. D. (2004). *Strangers to ourselves*. Harvard University Press.

Young, S. (2016). What is mindfulness? a contemplative perspective. In *Handbook of*

*mindfulness in education* (pp. 29–45). Springer.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, *9*(2p2), 1.