Erkenntnis
https://doi.org/10.1007/s10670-023-00669-4

ORIGINAL RESEARCH

# A Legion of Lesions: The Neuroscientific Rout of Higher-Order Thought Theory

Benjamin Kozuch[1]

Received: 5 April 2022 / Accepted: 12 February 2023

**Abstract**
Higher-order thought (HOT) theory says that a mental state is conscious when and only when represented by a conceptual, belief-like mental state. Plausibly, HOT theory predicts the impairment of HOT-producing brain areas to cause significant deficits in consciousness. This means that HOT theory can be refuted by identifying those brain areas that are candidates for producing HOTs, then showing that damage to these areas never produces the expected deficits of consciousness. Building this refutation is a work-in-progress, with several key components of it already existing in the literature (e.g., Boly et al. in J Neurosci 37(40):9603–9613, 2017; Kozuch in Mind Lang, 37(5):790–813, 2022). This article pulls together and supplements these earlier components so as to provide a comprehensive lesion-based argument against HOT theory, one which offers extra data and arguments in support of it, and which also addresses common objections.

## 1 Introduction

How far down the evolutionary and developmental scale does consciousness go? Are dogs conscious? Human infants? If one witnesses a dog's joy on being reunited with its person, or hears a baby's effervescent giggle, one certainly gets the impression of consciousness. But perhaps this is too quick: Whether any creature is conscious depends on what the cognitive requirements for consciousness are, and—according to some contemporary theories of consciousness—what is required for conscious states could be cognitively demanding indeed. According to *higher-order thought theory* (HOT theory), for example, consciousness is constituted by a certain kind of self-awareness, one that involves representing not just one's own thoughts and perceptions, but also using concepts to do so (Brown, 2015; Carruthers, 2000; Gennaro, 1996; LeDoux & Brown, 2017; Rosenthal, 2002). But it is plausible that numerous mammals, including babies and dogs, lack the cognitive sophistication

✉ Benjamin Kozuch
bkozuch@ua.edu

[1] Department of Philosophy, University of Alabama, Tuscaloosa, USA

Published online: 23 March 2023

required for this, in which case HOT theory implies that such creatures lack consciousness.[1] Thus, the debate of over HOT theory is far from academic: Its outcome might very well change which creatures we take to be capable of (conscious) pain and pleasure, and this in turn might change which beings we consider worthy of moral consideration.[2]

It has been difficult to determine whether HOT theory is true or not, and the numerous a priori arguments offered for and against it have arguably led to stalemate.[3] This makes more welcome the recent attempts to use neuroscience to shed light on the debate (e.g., Lau & Brown, 2019; Sebastián, 2014). Notably, neuroscience presents an opportunity to refute HOT theory (Kozuch, 2014, 2021): If HOT theory is true, there must be one or more brain areas (or networks of areas)[4] such that, because they produce HOTs, their functioning properly is necessary for consciousness. Since impairment of these areas would be expected to create conscious deficits, this provides a way to present a particularly strong argument against HOT theory: One starts by comprehensively identifying those brain areas that are candidates for producing HOTs, then goes on to show that damage to them never produces the kinds of conscious deficit to be expected when integral areas are damaged.

Note now that important pieces of this refutation are already in place: It is often hypothesized that the prefrontal cortex (PFC) is a particularly good candidate for producing HOTs. However, recent surveys of PFC lesion data show damage there to not produce deficits in visual experience (Boly et al., 2017; Kozuch, 2014), not of the type expected when integral areas are damaged (Kozuch, 2021).[5] Nonetheless, the case against HOT theory is incomplete, since a review of the debate reveals outstanding issues, including whether neuroplasticity could mitigate the effects of integral area damage (Lau & Rosenthal, 2011; Morales & Lau, 2020); whether PFC damage fails to produce conscious deficits because there are non-PFC integral areas (Gennaro, 2012); and whether damage to a HOT-producing mechanism would produce metacognitive deficits that prevented subjects from reporting their conscious deficits. This article addresses these and other relevant issues, doing so while presenting an updated and more comprehensive case for the idea that available brain lesion data count strongly against HOT theory.

Before beginning, I will lay out the argument more precisely, which looks as follows: If HOT theory is true, then there are one or more *integral areas*: These are

[1] I take it that it is currently *at least* an open question as to whether HOT theory implies that less sophisticated creatures are not conscious, in which case the issue of whether HOT theory is true or not remains important. For arguments that HOT theory entails that simpler beings are not conscious, see (Carruthers, 1992; Dretske, 1995); for objections to this, see (Brown et al., 2019; Gennaro, 1993; Lau & Rosenthal, 2011); for criticisms of some of these objections, see (Carruthers, 2000).

[2] Carruthers argues that there is a way to view some non-conscious animals as nonetheless deserving of moral consideration (1999).

[3] For an illuminating exchange that is recent, see (Block, 2011; Rosenthal, 2011; Weisberg, 2011).

[4] The possibility of there being more than one brain area involved in the production of HOTs is discussed below (Sect. 2).

[5] For a similar line of argument recently offered, see (Raccah et al., 2021), where it is argued that direct stimulation of the PFC does not produce the kinds of effect on consciousness that would be expected if the PFC were essential to consciousness.

brain areas (or networks of areas) such that, because they constitute a HOT-producing mechanism, their functioning properly is necessary for consciousness. So, given the way that "integral areas" are defined, they exist if and only if HOT theory is true.[6] Thus, this article argues against HOT theory by presenting evidence against there being any integral areas. The plan is to first identify all of those brain areas that are good candidates for producing HOTs (what we can call "potential integral areas"), then to show that none of them appear to be an integral area, since damage to each of them never produces the kinds of deficit in consciousness to be expected in cases of integral area damage. Like much research in this area, the investigation will be limited to *visual* consciousness, this being the area for which we have the richest data available. And so what I argue for here, more precisely, is that there are no visual integral areas, i.e., no HOT-producing areas whose functioning properly is necessary for having visually conscious states.

The article is structured around the requirements for arguing that there are no (visual) integral areas: Sect. 2 establishes where the potential integral areas are located; Sect. 3 explains what kinds of conscious deficit should result when integral areas are impaired; Sect. 4 shows that the lesion data concerning the potential integral areas presents no evidence of the predicted deficits; Sect. 5 argues against the idea that the failure to find deficits can be explained by claiming that integral area damage typically does not result in its impairment (because of, e.g., neuroplasticity); Sect. 6 answers objections.

## 2 The Location of the Potential Integral Areas

This article argues against HOT theory by showing that there are no integral areas. It does this by showing that damage to those brain areas that are candidates for producing HOTs never produces the expected deficits in visual consciousness. The question arises as to which brain areas must be considered in this investigation, that is, what we should take to be our *potential integral areas*. That is the question that this section addresses. We start with an explanation as to what HOTs are.

A HOT is a conceptual, belief-like state that represents a mental state as being a mental state of the subject (Gennaro, 1996:58; Carruthers, 2000:227; Rosenthal, 2002:410).[7]According to HOT theory, for instance, it is only when the subject has a higher-order representation with the content “I am seeing red right now” that one has a conscious representation of redness. Something to be pointed out right away

[6] And if HOT theory is false, then any HOT-producing mechanisms that exist turn out not to be integral areas; to mark the distinction, we can say that these areas would be *mere* HOT-producing mechanisms.

[7] The references here represent the locus classicus of HOT theory, but many newer higher-order (HO) theorists have views that are essentially the same, in that they hold (or appear to hold) that the right kind of HO representation will need to (1) be conceptual, and (2) represent mental states as being states of the subject (Brown 2015; Ledoux & Brown 2017; see also Brown, Lau, and Ledoux 2019); however, other HO theorists reject these requirements (Cleeremans, 2011; Lau, 2019). These latter theories do not technically count as forms of HOT theory, as this article understands it, but that does not mean that they are immune to the argument given here; this is a matter to which we will return in the article’s conclusion.

is that HOTs are *sophisticated* and *compound*, in that they involve representing not only a mental state and oneself, but also representing the mental state *as* a state of oneself, doing all this using concepts. So, the bar appears high when it comes to qualifying as a potential integral area, not only because an area must produce states of this sophisticated, compound sort, but also that it must make enough of them (or complex enough of them)[8] to be sufficient for supplying all of the conscious states composing our ongoing visual experience, this probably being a large amount (Carruthers, 2000, Chap. 8; Kozuch, 2018, 2021; Siewert, 1998, Chap. 7). It would not be surprising if not many areas qualify as potential integral areas.

The issue of what the potential integral areas are has already been addressed to some extent, with a near-consensus emerging that the best candidates for producing higher-order (HO) states are in the PFC (Kriegel, 2007, Block, 2009, Lau, 2011, Lau & Rosenthal, 2011, Kozuch, 2014, 2021, Sebastián, 2014, Lau & Brown, 2019). This view is largely motivated by how the PFC carries out advanced, often metacognitive types of function, such as working memory, action-monitoring, or self-awareness (Miyake et al., 2000), i.e., those functions that might plausibly involve producing HOTs. The dorsolateral PFC (dlPFC) is considered a particularly strong candidate, both because of studies showing apparent correlations between dlPFC activity and visual consciousness (for review: Lau & Rosenthal, 2011), and because of its involvement in perceptual metacognition (e.g., Chiang et al., 2014; Miyamoto et al., 2017; Rounis et al., 2010). There are, as well, strong theoretical reasons for thinking that dlPFC creates not just HO states (of one kind or another), but specifically HOTs: One of the functions that dlPFC carries out is conceptual short-term memory (Owen, 1997; Potter, 1999), this being the cognitive function that has been determined to be the one most likely to involve the production of HOTs (Beeckmans, 2007). This last observation, i.e., the idea that dlPFC is probably a HOT-producing area, is particularly important to note, since it means that, if we find instances where dlPFC has been impaired, they can plausibly also be taken to be instances where *specifically a HOT-producing mechanism* has been impaired. We return to this idea below.

In any case, this article also adopts the popular idea that the PFC contains potential integral areas, the dlPFC being an especially good candidate. However, the question still remains as to whether there are areas outside of the PFC that are potential integral areas. As we see now, there are a few.

First, since HOTs are a kind of *self-awareness*,[9] those brain areas that are involved in this function should be included in the potential integral areas. Here, there are two prominent theories. The first is Northoff and Bermphol's theory that "self-referential stimuli" are processed in the "cortical midline structures," these being the orbitofrontal and ventromedial PFC, and the anterior and posterior cingulate cortices

[8] The parenthetical phrase here recognizes certain options available when hypothesizing what kind of, and how many HOTs construct our ongoing conscious experience: It could be that there are numerous HOTs, each with very specific and limited contents; or it could be just a few HOTs with more complex contents; or perhaps there is just one HOT with very complex contents.

[9] Insofar as HOTs involve attributing mental states to oneself.

(Northoff & Bermpohl, 2004). The second is Damasio's theory of the "core self," something said to arise when internal and external stimuli are conjoined in two brain areas, the ventromedial PFC and posterior cingulate cortex (Damasio, 1999). Note now that both theories include a non-PFC area, the posterior cingulate cortex, which we can now add to our list of potential integral areas.[10] Another function possibly involving self-awareness would be visual metacognition (e.g., the ability to judge how confident one is about their perception), a function mostly carried out in the PFC, but which does involve a non-PFC area, the *insula* (an area deep in the brain's lateral fissure) (Vaccaro & Fleming, 2018). Other areas worth considering here are those carrying out *theory of mind*, since one version of HOT theory (Carruthers, 2000) hypothesizes that the mind-reading mechanism is what produces HOTs. The theory of mind mechanism includes areas not just in the PFC (medial PFC) (Bird et al., 2004), but also the temporoparietal junction (Samson et al., 2004).

We have found three non-PFC potential integral areas: the posterior parietal cortex, temporoparietal junction, and insula. Naturally, something we lacked space for here is going through the entire rest of the brain, saying why each area should not be considered potentially integral. Given this, we cannot rule out successful arguments being made for there being more potential integral areas than have been considered here. However, the method by which the relevant areas were identified has been conscientious if not extended,[11] and my hope is that it is at least the case that the burden of proof has been shifted to whoever would claim that the list of potential integral areas that we are using is incomplete. In any case, it is probably not *crucial* that lesion-based arguments against HOT theory even consider areas outside of the PFC, for reasons that I explain now.

We start by asking the question as to whether it is possible for HOTs to be produced by some area, or network of areas, that are located wholly outside of the PFC. This seems unlikely since HOTs are a sophisticated and complex kind of mental state, one that involves the integration of multiple other mental states. This all sounds like the province of the PFC (Kozuch, 2014; Kriegel 2007; Lau 2010; Metcalfe & Schwartz, 2015; Miyake et al., 2000). Thus it is not surprising that, in the case of each of the three extra potential integral areas that were found, the network that they were a part of also included areas in the PFC. Now, if all HOT-producing networks probably include the PFC, the following question arises: If just those components of the network that are in the PFC are impaired, should we expect this to cause deficits in HOTs? (As opposed to the network being able to carry on without

[10] The anterior cingulate cortex need not be included, as it is considered part of the PFC (Andrewes 2001, Chap. 3).

[11] The method was to start by generating a list of cognitive functions, one pitched at general enough of a level to allow it to hopefully be somewhat comprehensive. The functions included attention, learning, language use, long-term memory, perception, mind reading, thought and planning, self-awareness, and working memory. After making this list, I asked which of these functions might not only produce HOTs, but also produce enough of them to populate the entirety of our ongoing visual experience. It seemed that the only plausible candidates were working memory, self-awareness, mind reading, and attention. Three of these appeared in the discussion above. In the case of the remaining function, attention, it seems that, of the multiple types of attention there are, the one that would be relevant is *executive* attention (Zimmer, 2008), but this is a function that is primarily carried out by just the PFC and sensory areas.

the help of the PFC.) This question is important because, if damage to just the PFC areas creates deficits, then if it were shown that individually damaging each part of the PFC does not produce deficits in consciousness (something we do below), it would mean that there are probably no integral areas.

There are reasons for thinking that damaging just the PFC areas in a HOT-producing network would produce functional deficits. A first reason comes from the numerous cases in neuropsychology where damage to individual network nodes creates significant functional deficits. For some examples: Attentional deficits result from damage to the frontal-eye fields, intraparietal sulcus, temporo-parietal sulcus, *or* the middle frontal gyrus (Bartolomeo, 2021; Karnath, 2015; Ptak & Schnider, 2011; Toba et al., 2018); deficits in gaze shift result from damage to either areas in the posterior parietal cortex or in the dorsal premotor cortex (Ptak et al., 2017); and deficits in mind-wandering results from damage to the parietal lobule, the inferior frontal gyrus, or the dorsal, ventral, or anterior medial PFC (Philippi et al., 2021).[12] Consider now that, with HOT-producing networks in particular, there is even stronger reason to expect damage to individual nodes to cause functional deficits, since so much must be in place before a HOT can be successfully formed: A HOT requires concepts, representing mental states, representing oneself, and representing mental states as states of oneself. If a node responsible for supplying any *one* of these aspects of a HOT fails to play its part, then a HOT is not formed, and a conscious state is lost.

What this all means is that a strong case against HOT theory can be made by examining just the results of lesions to the PFC. Nonetheless, an even stronger case can be made by also showing that damage to non-PFC potential integral areas also fails to produce the expected deficits.[13] Given this, the three potential integral areas identified above are still included in the survey of lesion evidence carried out in Sect. 4.

## 3 Integral Area Impairment Would Cause Dramatic Deficits of Consciousness

In some cases of integral area damage, this might cause functionality to be lost, resulting in a degraded ability to produce HOTs. Let us mark this terminologically, by saying that, in those cases where an integral area is damaged, and it is experiencing functional deficits, the integral area is "impaired" (otherwise, it is "merely" damaged). While it is possible that damage to an integral area would not cause impairment (an issue discussed in Sect. 5), this section puts aside this possibility to investigate the question of how, in those cases where an integral area is impaired, we should expect it to change a subject's visual experience.

(This section leans heavily on (Kozuch, 2021), insofar as it largely takes for granted that the deficits following integral area impairment would be as they are

[12] This is the so-called "default mode network".

[13] It would further decrease the possibility that there could be integral areas outside of the PFC.

described in that article. Extended argumentation for this conclusion is to be found in the just-cited source. The goal here will instead be to just to describe the deficits, and to further highlight their striking nature.)

We start with some observations about visual experience. Consider how visual experience is composed of the representation of properties, ones such as brightness, color, motion, and shape (to name a few), with each represented property appearing in some part of one's visual field. According to HOT theory, each of these representations composing one's ongoing visual experience is conscious in virtue of its being represented by a HOT (Carruthers, 2000, Chap. 8).[14] So, in moments where an integral area's damage interferes with its ability to represent visual states, some of those visual states that would have composed the subject's ongoing visual experience (i.e., ones representing properties such as color, shape, etc.) might instead fail to be conscious, creating "gaps" in the subject's experience. This raises the question as to what it would be like for one to lose conscious states in this way.

It seems that we have but one existing model for this, one coming from the types of blindness and agnosia that result from visual system damage. For example, in cases where integral area damage caused a *complete* inability to produce HOTs, and therefore a complete loss of visually conscious states (perhaps an unlikely scenario), this would resemble cases where severe damage to the visual system caused a complete inability to produce visually conscious states, this being the condition that we refer to as "complete blindness." In cases of partial impairment of the ability to produce HOTs, the resulting deficits would be determined by which first-order visual states failed to be targeted by HOTs. For instance, if the impairment meant that all of the visual states representing properties in the *center* of the subject's visual field went unrepresented by HOTs, this would phenomenologically resemble a centrally located scotoma (i.e., a blindspot). Or it might be the case that the damage caused a certain type of *property* to fail to be the target of HOTs, in which case it would phenomenologically resemble a type of visual agnosia; for example, in the case where color representations failed to be targeted by HOTs, this would resemble acquired color blindness ("achromatopsia" (Zeki, 1990)), and the subject would be prevented from consciously experiencing color in some portion of their visual field (perhaps instead experiencing that portion in black-and-white)[15]; if motion representations failed to be targeted, this would resemble motion blindness ("akinetopsia" (Zihl et al., 1983)), causing an object's motion to appear like a series of static images rather than being smooth. Integral area impairment could as well cause the loss of other properties from visual experience, such as shape (Heider, 2000), identity (McCarthy & Warrington, 1986), or others.

Deficits that resemble scotomas or agnosias already sound phenomenologically remarkable. But consider now that these gaps in experience are likely to be

[14] It could be that the HOT in question represents only that visual state, or that it is one among many that the HOT represents.

[15] In at least some cases of achromatopsia, the subject does not describe the world as monochromatic, but rather just "responds in a way that suggests that color no longer makes sense as a term" (Robert Kentridge, personal communication). While intriguing, this empirical observation does not affect what would be predicted to happen in those cases where color representations fail to be targeted by HOTs.

distributed around the visual field and among types of visual experience. One reason to think this comes from the nature of the PFC (i.e., the part of the brain most often thought to contain potential integral areas (see Sect. 2)), since areas in the PFC appear to be undifferentiated geographically: Unlike most visual areas, areas in the PFC are rarely retinotopically organized,[16] and are also not separated into divisions that each specialize in processing a certain type of property (e.g., color or motion) (Jerde & Curtis, 2013). This lack of geographical differentiation, along with the stochastic nature of brain damage, means that deficits caused by integral area damage probably would not be localized to any one part of the visual field or type of visual experience, instead being distributed among them. Also contributing to this would be the way in which neurons in the PFC have *mixed-selectivity*, i.e., the ability to participate simultaneously in multiple representations (Fusi et al., 2016; Rigotti et al., 2013). That PFC neurons are representationally dynamic in this way would furthermore suggest that which first-order states failed to be represented would change over time. It seems, then, that we should expect the phenomenological scotomas/agnosias caused by integral area damage to be scattered throughout the visual field and between types of visual experience, and that the position and nature of these deficits would probably change over time. For example, the deficits might cause the subject to have a peripherally located scotoma at one moment, a lack of color experience in central vision in the next, then maybe a failure to experience an object's shape, and so on. As well, there might be, at any one given time, deficits of multiple types located in multiple parts of the visual field.

Probably, deficits like these need not become too severe before becoming phenomenologically remarkable: While the deficits would be particularly dramatic if the ability to represent lower-order states was drastically reduced (say, if the damage caused more than half of the subject's ongoing visual experience to be replaced with scotomas/agnosias), we would guess that such deficits would still be phenomenologically conspicuous even in cases of minor impairments, for instance, if it was just ten percent of one's ongoing visual experience that was subject to these shifting sets of scotomas and/or agnosias.[17]

One might wonder here whether we should assume that the loss of visually conscious states caused by integral area impairment would manifest in this particular way, i.e., as deficits that phenomenologically resembled types of blindness and agnosia. But it is not clear what other options there are, since there do not seem to be any other models for the loss of visually conscious states, at least not empirically

[16] Many visual brain areas are organized such that parts of the visual field next to one another are represented next to one another in the brain, this being particularly true in the case of visual areas that are lower in the cortical hierarchy (e.g., V1).

[17] One might suggest that the gaps being scattered and shifting might make them subtle enough to miss detection by the subject, or at least make them less likely to be detected than in cases where the deficits are continuously concentrated in one part of the visual field. But a potential for this to happen could probably only arise in those cases where the impairment was very mild; if it was not mild, the phenomenological gaps that were simultaneously shifting around the visual field would doubtless become a spectacle unto themselves. However, the evidence we review next will include instances where HOT-producing mechanisms appear to be significantly impaired, in which case we should expect even shifting and scattered deficits to be detectable by the subject.

based models (that is, ones found in the existing brain lesion literature).[18] Given this, we should think that integral area impairment would bring about deficits of the sort described above—at least until another plausible suggestion is produced.

To summarize, it seems that integral area impairment would cause the appearance of many "gaps" in the subject's visual experience, where these are portions of their experience that phenomenologically resembled scotomas and/or agnosias, and which might dynamically change and shift about the visual field. It is reasonable to guess that dramatic deficits like these would be phenomenologically striking.

Now would be a good time to note something else: The fact that integral area impairment would cause fairly dramatic deficits of consciousness means that the PFC lesion studies sometimes offered in support of HO theories (e.g., Barcelo et al., 2000; Del Cul et al., 2009; Rounis et al., 2010) do not actually confirm HOT theory,[19] but rather disconfirm it. Here is not the place for extended argumentation for this conclusion (for this, see Kozuch, 2021), but some familiarity with the idea is useful for what comes below.

Let us consider one of the studies in question: In (Rounis et al., 2010), subjects were presented with a stimulus (a square next to a circle) followed by a mask (two stars in the same locations). Transcranial magnetic stimulation (TMS) was applied to subjects' dlPFC before some trials (creating what could be considered "temporary lesions"), and this caused the subjects to (a) become worse at making metacognitive judgments (ones concerning how confident they were as to whether they saw the target stimulus), and (b) become more likely to indicate that they did not consciously perceive the stimulus. Since the metacognitive deficits that these subjects suffer from are plausibly construed as deficits in HOTs, these studies are taken to support HOT theory since they demonstrate correlations between deficits in HOTs and deficits in conscious states (e.g., Lau & Rosenthal, 2011). But what is important here is that whatever conscious deficits these subjects might be experiencing[20] do not remotely resemble those predicted in cases of integral area impairment; they consist, at best, just of a failure to consciously experience stimuli that are localized to a very small part of the visual field (i.e., where the stimuli appeared); this is a far cry from the kinds of widespread and non-localized conscious visual deficits to be expected in the case of integral area impairment. The same kinds of observation can be made about many of the other studies to which HOT theorists have appealed (Lee & D'Esposito, 2012; Philiastides et al., 2011; Turatto et al., 2004): In each case,

[18] As well, it is difficult to imagine what this would be like, outside of the empirical examples. Perhaps the degraded ability to produce HOTs might bring about some kind of "phenomenological fading," a loss of the vivacity of some or all of one's visually conscious states (e.g., perhaps the conscious states would take on something comparable to the phenomenologically faded appearance that memories have). However, imagined alternatives like this can help the HOT theorist only if the deficits are significantly less remarkable than the ones described above, but it seems like we would expect "phenomenological fading"—and probably any other imagined alternative—to be no less remarkable.

[19] For examples of these data being used to support HO theories and/or the idea that the PFC is essential for consciousness, see (Lau & Rosenthal, 2011; Odegaard et al., 2017).

[20] There is a strong case to be made for subjects in these experiments not having deficits of consciousness at all (see Kozuch 2021).

these studies either fail to demonstrate conscious deficits at all, or whatever conscious deficits they do demonstrate are too minuscule to count as being the kind to be expected in cases of integral area impairment.

It will be useful to refer back to the Rounis et al. study from time to time. For now, the intention has just been to point out how many of those studies that have been thought to support HOT theory turn out not to do so. And so the reader should not be surprised to see these studies later cited as evidence against HOT theory.

## 4 Damage to the Potential Integral Areas does not Produce the Expected Conscious Deficits

### 4.1 Review of Lesion Evidence

Having determined what the results of integral area impairment would be, and where the potential integral areas are located, we are now poised to review the lesion evidence, looking for the predicted deficits. The discussion is divided into PFC and non-PFC areas.

PFC damage tends to cause deficits in so-called *executive function* (Miyake et al., 2000), abilities such as reasoning, planning, or advanced metacognition; however, nowhere to be seen are the kinds of conscious visual deficits that would result in cases of integral area impairment (Kozuch, 2014, 2021; Pollen, 1995, 2007): Removal of the entire (or at least most)[21] of the PFC has been shown to cause a reversion to child-like behavior and intelligence (Brickner, 1936, 1952), and "massive bilateral prefrontal damage" has been shown to cause deficits in motivation and long-term planning (Markowitsch & Kessler, 2000); as well, unilateral ablations of the entire right or left PFC have been shown to lead to problems in multi-tasking or initiative (Penfield & Evans, 1935). However, nowhere in these studies is there evidence suggesting that integral area damage has occurred, since these subjects neither present with deficits in visual tasks,[22] nor report experiencing the kinds of gaps in visual experience that integral area impairment would cause. Indeed, in a review of two hundred PFC patients carried out by Feuchtwanger (1923), only two cases of perceptual deficits were found, bringing Feuchtwanger to conclude that PFC lesions bring about "psychic disturbances within the fields of emotion and performance, and not in the field of concrete actions,…[such as] perception, memory, thinking, movement, etc."[23] There are, as well, numerous other cases where PFC patients are examined or tested without the predicted defects of visual experience being found

[21] Whether or not the lesions in the Brickner study encompass the *entire* PFC is the subject of debate, with Odegaard et al. arguing that it is just most (2017), and Boly et al. arguing that it is all of it (2017).

[22] The issue of whether or not we should expect integral area impairment to cause deficits in detecting and/or reporting on a stimulus is discussed next, in 3.2.

[23] The Feuchtwanger quote is acquired from (Markowitsch 2000).

(e.g., Barcelo et al., 2000; de Graaf et al., 2011; del Cul et al., 2009; Fleming et al., 2014).[24]

So far, we have examined cases in which there is generalized unilateral or bilateral PFC damage, finding no evidence of the predicted deficits. The same is true in cases of unilateral or bilateral damage to any of the more specific parts of the PFC. According to one common way in which the PFC is divided up (Andrewes, 2001, Chap. 3), it includes four major divisions, these being the orbital, lateral, and medial PFC, along with the anterior cingulate cortex. Subjects with damage to just one of these areas might present with deficits in short-term memory (lateral PFC) (Thompson-Schill et al., 2002), theory of mind (medial PFC) (Bird et al., 2004), or social inhibition (orbitofrontal PFC) (Damasio, 1994). However, deficits in visual tasks or defects in visual experience are never reported in such studies. This is also true of a study in which twenty-three patients were each given a bilateral ablation isolated to one of these three areas (i.e., orbital, lateral, or medial PFC), and then were tested extensively for visual defects (Heath et al., 1949). In the case of dlPFC (i.e., our best candidate for being an integral area), lesions here have been shown to produce deficits in remembering a stimulus (Thompson-Schill et al., 2002) or in successfully associating words (Stone et al., 1998), but nowhere in the large number of studies concerning dlPFC lesions is there even anecdotal evidence of patients experiencing anything like what would result in cases of integral area damage (Chiang et al., 2014; Philiastides et al., 2011; Rounis et al., 2010; Rowe et al., 2001; Turatto et al., 2004; for review, see Kozuch, 2021).

In the case of damage to the anterior cingulate cortex (ACC), this can result in the subject becoming passive and unresponsive (Cairns et al., 1941), something that has been interpreted as a lack of consciousness (Damasio, 1994). However, given that one of the ACC's primary functions is motivation (Devinsky et al., 1995), the patients' unresponsiveness can be explained as resulting just from a lack of motivation (Boly et al., 2017; Kozuch, 2014), with there being no need to hypothesize a lack of consciousness. This interpretation is supported by reports from former akinetic mutes, who describe themselves as having been aware of their surroundings but lacking any urge to do anything (Damasio & van Hoesen, 1983). Another disorder to consider here, since it sometimes results from PFC damage (to the frontal eye-fields (FEF)), is *hemispatial neglect*, a condition which consists of an inability to notice items appearing in the contralesional half of one's visual field, usually the left (Husain & Kennard, 1996). However, since FEF is part of an attentional network (Bartolomeo et al., 2007), the patients' inability to notice left-located items can be entirely explained by the attentional deficits that the FEF damage causes (Brogaard, 2011a, 2011b; Jacob & de Vignemont, 2010; see esp. Kozuch, 2014, 2022), there again being no need to hypothesize a lack of consciousness. Something supporting this interpretation is the fact that bilateral damage does not produce an inability

[24] I remind the reader that, though some of the studies just cited have been touted as evidence *for* HOT theory, they act as evidence against it because the conscious deficits seen in these experiments (which are small or none) are not what would expected in cases of integral area impairment (for extended argument, see Kozuch 2021).

to notice items in *both* visual fields (Cazzoli et al., 2012; Pierrot-Deseilligny et al., 1986), where this is what should probably be expected if unilateral damage actually produced a lack of consciousness of the contralesional visual field (Kozuch 2022).[25]

Now we review data concerning lesions to potential integral areas outside of the PFC, these being the posterior cingulate cortex (PCC), temporoparietal junction (TPJ), and insula. Lesions to the PCC create deficits related to long-term memory: A subject with PCC damage might have difficulty in retaining details of stories (Kasahata et al., 1994; Katai et al., 1992; Valenstein et al., 1987), in remembering the relative spatial locations of objects (Takahashi et al., 1997), or in forming memories about events experienced by oneself (Gainotti et al., 1998). Lesions to the TPJ causes deficits in attention and theory of mind, which can manifest as a decreased susceptibility to distracting stimuli that is presented contralesionally (Ro et al., 1998; Samson et al., 2004), or as an impaired ability to infer false beliefs in characters in stories (Samson et al., 2004). However, in the case of both the PCC and TPJ, there is no evidence of the lesions having caused defects of consciousness. And while damage to the insula sometimes produces perceptual deficits, such as ones concerning audition or pain, there is no evidence for any of these deficits being visual (Bamiou et al., 2003).

There is something important to note about the lesion evidence that we have just reviewed, which is that it very likely contains instances in which specifically a HOT-producing mechanism has been impaired. This is because the evidence contains cases of dlPFC impairment, ones in which subjects suffered from metacognitive deficits (e.g., Fleming et al., 2014; Rounis et al., 2010). Given that dlPFC is our best candidate for producing HOTs (see Sect. 2), this gives us justification for thinking that these metacognitive deficits will be deficits in HOTs in particular. And so it seems that the above evidence contains instances in which we can be pretty confident that a HOT-producing mechanism has been impaired—and yet the dramatic deficits of consciousness are still nowhere to be found.

In summary: Our survey of lesion evidence concerning the potential integral areas turned up no evidence—not even anecdotal—of subjects suffering from the kinds of conscious deficit that come from integral area impairment. This is true even though the evidence appears to contain instances in which HOT-producing areas have actually been impaired. This seems to count strongly against HOT theory. It would, at least, if we could be sure that any dramatic deficits of consciousness would be likely to be detected. This is the topic of the next section.

[25] One might also attempt to support HOT theory by appealing to some recent studies concerning prefrontal seizures (Bonini et al., 2016), but the experimenters in this study appear to equate the ability to make reports with the ability to have conscious states, an assumption to which they are probably not entitled (see Block's discussion of petit mal seizures in his (1995)).

### 4.2 The Protocol Used to Examine Brain-Damaged Subjects is Likely to Find Any Dramatic Deficits of Consciousness

One might wonder whether the deficits of consciousness predicted by HOT theory would be detected in a clinical or experimental setting. As we see now, if we take a look at the protocol typically used to examine brain-damaged subjects, it seems very likely that any striking deficits of visual consciousness that the subjects suffered from would be discovered.

The treatment of brain-damaged patients usually starts with an interview meant to serve as a first pass at determining the lesion's effects (Hebben & Milberg, 2009, Chap. 3). Such an interview typically starts with an open-ended invitation to the patient to "describe whatever problems they are having" (ibid., p.61). It is of course unlikely that, at this point, the patient would not mention any significant deficits of consciousness like those that would result from integral area impairment. But even if we assume that a subject failed to mention them at this stage—perhaps the subject, improbably, had not yet noticed them—it is likely that they would notice and report on them during the battery of tests given next, since these frequently include not only tasks requiring visual perception as a component (e.g., copying drawings, arranging blocks in a frame, reading strings of digits), but also ones that explicitly test for normal vision (e.g., identifying colors or objects) (Cullum, 1998).[26] And even if patients make it this far without somehow revealing their deficits, some of the patients would have had further opportunities to do so, since many of the lesion studies discussed above utilized experiments that were designed specifically to gauge whether the subjects had defects in their visual abilities or experience (e.g., Heath et al., 1949; Turatto et al., 2004; Flemming et al., 2014).

There are further reasons to think that the deficits would be discovered. First, the conscious visual deficits might cause easily detectable *functional* deficits. One possibility here is that the nature of phenomenological blindspots is such that the lack of conscious perception in the blindspot prevents the subject from making accurate reports on what appears in that part of the visual field, in which case the presence of deficits in the subject's visual experience could be gleaned from their poor performance on visual tasks. However, it is sometimes argued that a loss of HOTs should not be expected to necessarily prevent a subject from reporting on stimuli, as long as the relevant first-order states still obtained (the first-order states would just fail to be conscious) (Lau & Rosenthal, 2011). But this makes the deficits no less likely to be discovered: If the subjects could report on stimuli that they did not consciously experience, this would subjectively resemble "blindsight," the phenomenon in which subjects with a damaged primary visual cortex can accurately report on items in parts of the visual field to which they subjectively seem blind. Blindsighters are usually surprised to discover that they have this ability (Weiskrantz, 1986, Chap. 2), and we would guess that subjects with integral area damage would be no less surprised when finding out that they could (easily) visually detect items that they failed to

[26] Tasks such as these are included as part of the commonly used Halstead-Reitan Neuropsychological Battery (Reitan & Wolfson, 1985).

consciously see.[27] And, of course, such an ability would also probably come to light during neuropsychological exams.

There seems to be a strong prima facie case in favor of the idea that, if some of the subjects in the lesion data surveyed above had the predicted dramatic deficits, there would be evidence of them. However, the HOT theorist might argue that, though the evidence reviewed contained instances of integral area damage, such deficits failed to appear because the damage did not cause the integral area to become impaired. We consider this possibility next.

## 5 Neuroplasticity Would not Prevent the Conscious Deficits

It is often argued that, if an integral area is damaged, it is possible that neuroplasticity (the brain's ability to adapt to damage or experience) would prevent the appearance of deficits in visual experience (e.g., Kriegel 2007; Lau & Rosenthal, 2011; but see Kozuch, 2014, 2021).[28] On its face, this objection looks weak, given that neuroplasticity rarely provides full recovery, especially in adults (Grafman, 2000; Grafman et al., 2010), with it sometimes even having been argued that "true" recovery never happens (Krakauer, 2006; Nudo, 2011).[29] Moreover, there have already been forceful arguments made against the idea that neuroplasticity would prevent the appearance of conscious deficits in cases of integral area damage (Kozuch, 2014, 2021). Nonetheless, commentators continue to maintain that this is possible, with it being especially likely in the case of the PFC (Lau & Rosenthal, 2011; Morales & Lau, 2020; Odegaard et al., 2017), a contention perhaps made more plausible by how PFC neurons are more representationally dynamic than those in other parts of the brain (the "mixed selectivity" mentioned above). But the evidence in favor of strong PFC neuroplasticity is thin.

Often, when commentators defend the idea that the PFC is especially resilient against damage (e.g., Lau & Rosenthal, 2011; Odegaard et al., 2017), the only study cited in support of this is one by Voytek and colleagues (2010). In this experiment, subjects with unilateral PFC damage were shown to have increased activation

[27] Note that we would expect the blindsight-type abilities that such subjects displayed to go well beyond what is seen in the experiments to which HOT theorists sometimes appeal (Lau & Passingham, 2006), in that such subjects would probably be able to do more than simply make accurate forced choices as to whether the unexperienced stimuli had been presented or not; they would probably also, for example, be able to do things like "free-report" its color or alphanumeric identity.

[28] A related objection here is one saying that such deficits are prevented by there being more than one integral area (Kriegel 2007); a detailed response to this is found in (Kozuch 2014:736–38). The general thrust of the response is that it is evolutionarily implausible that our cognitive system would expend the resources necessary to not only higher-order represent all of the lower-order states composing our ongoing visual experience (something already questionable (Carruthers 2000:213–22)), but to also do so redundantly.

[29] According to this skeptical view, in those rare cases where neuroplasticity appears to provide something close to full recovery (in an adult), this is always just because the subject now relies on different cognitive resources to accomplish the task (a phenomenon known as "compensatory masquerade") (Grafman & Litvan, 1999).

in their contralesional PFC on trials where they completed the task successfully. Voytek et al. hypothesized that the increased activation occurred because contralesional PFC was enlisted to make up for deficits caused by the lesion, and some commentators have taken these results to show that in cases of "localized unilateral lesions to PFC…other parts of the frontoparietal network can dynamically reorganize to compensate functionally" (Odegaard et al., 2017:9596). However, if these studies count as evidence for neuroplasticity preventing significant conscious visual deficits, then it needs to be not only the case that undamaged areas "*contribute* to functions normally carried out by the damaged side" (Lau & Rosenthal, 2011:369, italics mine), but rather that these contributions bring those functions up to something like a *normal* level; since if recovery were not complete, significant deficits in visual experience would still occur. But the experiment in question cannot present evidence for PFC neuroplasticity having brought about a normal level of function, since not only is it the case that the level at which the lesioned subjects performed was not measured,[30] but also that no normal level of performance was established at all (the experiment employed no controls).

There is, however, one other study that HO theorists (Lew & Lau, 2017) offer in support of strong PFC neuroplasticity: In (Mackey et al., 2016), subjects carried out a visual memory task, one in which they were asked to visually fixate the former location of a dot a few seconds after its removal. While subjects with precentral sulcus damage (an area in the PFC) showed significant impairment on this "memory guided saccade" task, subjects with dlPFC damage were unimpaired. Lew and Lau take this study to show that neuroplasticity in the PFC is exceptionally strong. Their reasoning is that, since dlPFC is closely associated with working memory, the lack of deficits resulting from dlPFC damage must be explained by neuroplasticity having made up for them. However, as the study's authors point out (pp. 2853–55), this explanation is problematic, since it is hard to explain why the strong PFC neuroplasticity hypothesized to eliminate deficits in dlPFC fail to do so when an adjacent PFC area, i.e., the precentral sulcus, is damaged. A better explanation is the one for which Mackey et al. argue, which is that, while human dlPFC plays a role in some kinds of working memory, it does not do so in the case of the memory-guided saccades used in this task.[31],[32], [33]

So, the studies to which HOT theorists appeal to support strong PFC plasticity fail to do so. But we should investigate whether there are any other lesion data that might accomplish this. One potential source comes from data concerning hemispatial neglect in monkeys (the disorder mentioned above, in which the subject fails

[30] Or at least not reported in the article.

[31] Thus their article's title, "Human dorsolateral prefrontal cortex is not necessary for spatial working memory".

[32] Notably, this explanation is consistent with prior dlPFC lesion data showing these deficits to be correlated with dlPFC damage, since all earlier studies using memory-guided saccades involved subjects that were either non-human, or whose damage included the precentral sulcus (Mackey et al., 2016).

[33] There are other studies that Morales and Lau (2020) appear to present as providing support for PFC neuroplasticity (Andersen et al., 1985; Barbas & Mesulam, 1981) but they do not appear to be relevant to this issue at hand.

to notice left-located items). In these studies (Adam et al., 2019, 2020; Rizzolatti et al., 1983), hemispatial neglect caused by targeted unilateral ablations to the PFC[34] would often resolve within six months (similar kinds of recovery from neglect is sometimes seen in humans (Ramsey et al., 2016)). However, the mere fact that neuroplasticity resolved these subjects' neglect cannot be taken to suggest that they do not suffer from other severe cognitive deficits (ones that plasticity failed to make up for), since the tasks used to test for neglect are *extremely* simple; a subject might be asked, for example, to simply detect whether an unobscured and long-lasting target appears in their left visual field. This leaves wide-open the possibility that these subjects would show severe deficits if given other, more complex tasks. Indeed, we should take it to be likely that they would, given the kinds of long-lasting and severe cognitive deficit that PFC lesions often produce (see, e.g., Szczepanski & Knight, 2014, and Sect. 4). Additionally, there are other recent monkey studies of prefrontal neglect caused by unilateral PFC damage in which there was no such resolution either after six months (Schiller & Chou, 1998) or a year (Schiller & Chou, 2000), and a similar lack of recovery from neglect is also often seen in humans (Farne et al., 2004; Ramsey et al., 2016; Rengachary et al., 2011). Perhaps more promising for HOT theory here is a study by Ainsworth and colleagues (2018), one in which two monkeys were given unilateral PFC lesions (in one case, to dlPFC) that resulted in deficits in working memory and object detection, ones resolving after three months. However, while this study (finally) gives us *some* evidence of a stronger form of PFC neuroplasticity, it only does so in the case of unilateral damage: After the first round of testing, the lesions were made bilateral, with this causing more acute deficits, ones still not resolved six months later.

I fear, however, that our extended investigation of neuroplasticity might have us losing sight of what is crucial here, which is not whether PFC neuroplasticity in adults is *ever* strong enough to bring about full recovery of some function, but rather whether we can count on these factors doing this *always or nearly always*; otherwise, we should expect to see remarkable deficits appear in visual consciousness in at least *some* cases of severe PFC damage, if HOT theory is true. But there are *so* many instances in which PFC damage creates significant and long-lasting cognitive deficits (Szczepanski & Knight, 2014), including ones in target detection (Barcelo et al., 2000; Yago et al., 2004), goal management (Goel & Grafman, 1995), and—in the case of dlPFC damage—working memory (Perlstein et al., 2004; Sanchez-Carrion et al., 2008; Serino et al., 2006; Voytek & Knight, 2010).[35] This being the case, we can be all but sure that PFC neuroplasticity does not always (or even nearly always) provide full recovery of function. Notice, moreover, that in the cases surveyed above, the lesioned subjects frequently suffered from severe and long-lasting deficits of one kind or another. This, of course, acts as further evidence against neuroplasticity being powerful enough to prevent deficits. But it also gives reason to think that, in any studies examined above in which an area containing a

[34] Lesions were often in the FEF (frontal eye fields), and sometimes included dlPFC.

[35] In the studies just cited, the lesions occurred anywhere between 6 months and 25 years prior to examination.

HOT-producing mechanism was damaged, the HOT-producing mechanism itself was impaired: It would be unexpected if damage to an area where a HOT-producing mechanism was located would create deficits in some of the functions that the area carried out without affecting others, especially in the case of what is probably one of the more demanding functions, that of HOT production. But we need not rely solely on such theoretical considerations to infer that the above data contain instances in which a HOT-producing mechanism has been impaired, since this was already established above (in Sect. 4.1), in the discussion of dlPFC studies, where it was found that some of these studies very likely present instances of this.

In this section, we have reconsidered the debate concerning whether neuroplasticity might prevent deficits in consciousness when an integral area is damaged, finding this idea to still look implausible.[36,37] However, there are still a couple ways in which the HOT theorist might object to the idea that the above data show that there are no integral areas. It is to these objections that we now turn.

## 6 Objections

Now we consider two objections, both of which are based around the way in which HOTs and reportability might be connected.

### 6.1 Objection 1: The lesion evidence includes no instances of HOT-producing mechanisms being impaired

The first objection goes as follows: The argument presented in this article disconfirms HOT theory by showing that damage to HOT-producing mechanisms never causes the predicted deficits of consciousness. But knowing that a subject has an undisturbed conscious experience requires the subject to report this, and this is something that plausibly might require HOTs. If this were true, then it seems that the studies reviewed above did not include instances in which subjects had an impaired HOT-producing mechanism, in which case these studies cannot count as instances of HOT-producing mechanisms being damaged without the predicted deficits occurring.

Let us evaluate this objection. A good place to start is by observing that using such an objection creates a debt for the HOT theorist, which is to explain why, if there are integral areas, damage to the potential integral areas never produces deficits in HOT production. What are the options here?

[36] A related objection here would be the idea that damage to integral areas might not cause conscious deficits because lower-order representations are often represented by HOTs *redundantly*. This issue is addressed extensively in (Kozuch 2014:736–738).

[37] The HOT theorist might raise the possibility that the brain would "fill-in" the gaps in visual experience that would be created by integral area impairment, especially if they were more scattered than concentrated. But this does not help the HOT theorist, since phenomenologically filling-in a gap requires no less HOTs than would be needed to prevent the gap in the first place, and it is already built into the case that we are considering that such HOTs are absent because of the damage.

One possibility is to claim that the above lesion evidence neglected to include the integral areas in the survey of lesion evidence. But this looks quite unlikely, given Sect. 2, where we addressed the issue of what the probable candidates are for being HOT-producing areas. Another way to explain the failure to find deficits is to hypothesize that the above data *did* include instances of integral area damage, but this failed to result in impairment; however, for this to be the case, neuroplasticity would have had to prevent, in all or nearly all cases, the damage from leading to impairment, another assumption that was above shown to be highly unlikely, this time in Sect. 5.

It seems, then, that this objection relies on one of two scenarios obtaining, both of which are suspect. Now notice that we have reason to think that one of the objection's premises is false: The objection claims that, in each of the studies above, the HOT-producing mechanism remains unimpaired, but this is belied by the dlPFC lesion studies in which subjects are plausibly construed as suffering from deficits in HOTs (see 3.1); as noted above, that these subjects suffer from such deficits seems to be an idea endorsed even by advocates of HOT theory (e.g., Lau & Rosenthal, 2011).

There seems, then, to be many things working against the first objection. Of course, something that has not yet been explained—but which must be explained by one who would use the lesion evidence to argue against HOT theory—is why having a damaged HOT-producing mechanism does not seem to interfere with subjects' ability to report on their experiences. A better time to address this appears below, after considering the second objection.

### 6.2 Objection Two: Subjects Cannot Notice the Deficits of Consciousness

The HOT theorist might agree that some subjects in the experiments above *do* suffer from dramatic conscious deficits, but argue that they are yet to be discovered. This could happen if the subjects themselves are unaware of the scotomas/agnosias. Precedence for this might come from the forms of *anosognosia* that exist, i.e., cases where brain-damaged patients remain unaware of deficits (e.g., that they have Alzheimer's, or that half of their body is paralyzed) (Vuilleumier, 2004). However, what is more important here is not anosognosia in general, but rather anosognosia for *visual deficits*, and, while it is true that scotomas caused by occipital damage are not always noticed by patients right away (e.g., Bisiach et al., 1986), this probably happens more rarely than researchers have thought it to in the past (Baier et al., 2015). In any case, what is important here is not whether subjects *sometimes* fail to notice scotomas, but whether they typically—sooner or later—*do* notice them, and the just-cited literature reveals that this is indeed the case.

However, the HOT theorist might argue that what we are considering here is a special case, doing so by appealing to the observation made in the last objection, which is that introspection might require HOTs. Here, it could be argued that, because HOTs and introspection might be closely connected, integral area impairment could lead to an especially strong form of anosognosia, one preventing the subject from discovering their deficits.

The important question here is whether an impaired ability to form HOTs would always (or nearly always) prevent the subject from noticing and reporting on the deficits. One way this could happen is if the impairment *always* caused a *complete* inability to introspect. This is unlikely, since the functional deficits caused by lesions are typically not total, this being something that can be gleaned from the survey of lesion patients carried out above (since these patients usually had functional deficits that were less than total).[38] Moreover, it seems like a disorder that consisted of a *complete* lack of ability to introspect would be clinically obvious—the subject would not be able to produce *any* reports about their mental life (or at least any accurate ones)—in which case we would expect to have found at least anecdotal evidence of such disorders in the evidence above.[39] Moreover, even if it were the case that damage to a HOT-producing mechanism *always* caused a complete inability to form HOTs, a residual ability to introspect would likely remain anyway, since it is commonly thought in cognitive science that there is probably more than one introspective mechanism (Hill, 2011; Prinz, 2004; Schwitzgebel, 2012).

But perhaps what is more important here is that there already seem to be *actual* cases where subjects retain some ability to introspect following impairment of their HOT-producing mechanism: Remember the Rounis et al. experiment (Sect. 2), the one where application of TMS to dlPFC caused subjects to be less likely to consciously experience the target stimuli. The reason that HOT theorists have in the past thought that this study supported HOT theory came from a correlation between deficits in HOTs and conscious states: That there is a deficit in HOTs was inferred from the subjects' metacognitive deficits, and that there is a lack of conscious states was inferred from the *subjects reporting* that they failed to experience the stimulus. So, if we assume that the reports of subjects in this experiment are reliable (as has often been done by HOT theorists), it seems that this study (and some of the others to which HOT theorists have appealed (e.g., Chiang et al., 2014; Fleming et al., 2014; Miyamoto et al., 2017)) present instances in which a subject has an impaired HOT-producing mechanism, and yet retain the ability to report on their mental states.

Notice now that these studies *also* give reason to think that this residual ability to introspect would be sufficient for being able to notice and report on whatever dramatic conscious deficits the subjects were suffering from: Subjects in these experiments, despite probably having an impaired HOT-producing mechanism, seem to have retained the ability to report the *absence* of conscious states (the ones that would have represented the target stimuli; i.e., the circle and the square). And if they

[38] For example, the PFC damage-caused deficits in short-term memory (Thompson-Schill et al., 2002) and theory of mind (Bird et al., 2004) that were discussed above were not total (i.e., the patients could still perform some tasks requiring these abilities; they were just significantly less accurate than the controls).

[39] Here, what is relevant are cases where the subjects' deficits are completely or mostly localized to HOT production, rather than ones where extensive brain damage (Odegaard et al., 2017) or intense intracranial stimulation (Quraishi et al., 2017) has brought about a generalized kind of cognitive incapacitation, one causing the subject to become catatonic or otherwise unresponsive. These cases are irrelevant because the subjects lack the ability to make any reports, including ones concerning their conscious experience, which means that we lack grounds for taking these studies to count either for or against HOT theory (Kozuch, 2014, 2021).

have the ability to notice small defects of consciousness like the one found in this experiment (where the defect consists only of a failure to consciously perceive the stimulus, and nothing else), we would guess that the subjects also have the ability to—at least, sooner or later—notice substantial defects of consciousness, like those created by integral area impairment. Indeed, it seems fair to make a more general point here, one extending beyond just the subjects in the dlPFC studies: Consider again that the integral area impairment would probably cause gaps in the subject's experience, ones that phenomenologically resembled scotomas and agnosias, are probably numerous, and might dynamically change and shift about the visual field. Given that these deficits would likely be a permanent feature of the subject's experience, the odds seem high that any subject that retained any ability to introspect would be able to detect the deficits—at least, sooner or later.[40]

Overall, there is good reason to think that, in the case of subjects in the above studies that had an impaired HOT-producing mechanism, there are instances in which the damage to the HOT-producing mechanism leaves at least a partial ability to introspect, and that this would allow the subject to notice any dramatic deficits of consciousness that they possessed. But consider now that, even if this was not possible, we already have reason to think that the deficits would be nonetheless discovered (see Sect. 4.2), since the impairment would either cause the subject to (a) have deficits in reporting stimuli (i.e., they would suffer from an inability to report on those stimuli that failed to be consciously represented), or to (b) exhibit blindsight-type symptoms, ones in which they were able to report on stimuli but not able to say how they did this (this ability would be sure to be remarkable to the subject and the experimenter alike).

In sum, there is good reason to think that the scenario described in the objection that we are considering, one in which damage to a HOT-producing mechanism prevents the subject from being able to notice the dramatic deficits of consciousness, is unlikely to obtain; and, even if it did, the deficits would probably be nonetheless detectable in a clinical or experimental setting.

What needs noted now is that, if it is true—as just argued—that a partial ability to introspect would be sufficient for making judgments about one's experience, this provides resources for giving an even stronger response to Objection One: The response given above consisted of a number of strong considerations against the idea that the lesion data surveyed above failed to contain instances in which a HOT-producing mechanism was impaired. This, however, left unanswered the question as how to explain the ability of subjects with a damaged HOT-producing mechanism to

[40] A question that we lack space for addressing in a detailed fashion is that of how exactly the introspective deficits would manifest. One possibility here is that the impairment would limit *which* mental states can be accessed (e.g., it might affect one's ability to access auditory but not visual states), but the representational flexibility that the PFC possesses makes this improbable (see Sect. 3). More likely, the deficits would affect the *reliability* and *accuracy* with which one can introspect; i.e., there might be occasions where the subject misidentified or failed to access a mental state, one that they wouldn't have misidentified or failed to access prior to the impairment. However, we would guess that as long as a minimal ability to introspect is maintained, the subject's introspection would be reliable/accurate enough to be able to eventually correctly identify the presence of dramatic deficits in their experience.

report on their experiences. And now we have seen that this is probably enabled by a residual ability to introspect.

Let us take stock: Speaking generally, there are three ways for the HOT theorist to explain why the investigation above failed to find the predicted deficits: First, the HOT theorist might claim that there are HOT-producing mechanisms located in parts of the brain outside of those considered in this article, but this looks implausible because it is very unlikely that any areas (or networks) wholly outside of the PFC could produce sophisticated states like HOTs (see Sect. 2). Second, the HOT theorist might claim that the lesion data did contain instances in which a HOT-producing mechanism was damaged, it is just that the damage failed to result in impairment, but this relies on a number of unlikely assumptions, the largest among these being that neuroplasticity is much more potent than we currently have reason to believe (Sect. 5). Third, the HOT theorist might claim that the above data contain instances where subjects do have conscious deficits (because they have a damaged integral area), it is just that it is difficult or impossible for the subjects to know about them (because of an impaired ability to produce HOTs), but this claim must assume that (a) a partial ability to introspect would not be enough to allow subjects to discover their deficits, even in the long term, and (b) the deficits would be clinically undetectable; however, we have just seen reason for doubting both of these ideas.

So, it seems that maintaining HOT theory in the face of the lesion data requires making at least one of three assumptions, each of which seems very unlikely, given the investigation above. On the other hand, adopting the idea that the lesion data disprove HOT theory involves only the assumption that a partial ability to introspect would be enough to detect striking, dramatic deficits of consciousness, or that the deficits would be otherwise detected by experimenters (because of, e.g., blindsight-like symptoms). This appears to be the most innocent of the assumptions that we have considered, and so the lesion data considered above are best explained by there being no integral areas, in which case HOT theory is false.[41]

## 7 Conclusion

At the beginning of the article, we saw that HOT theory seems to have to predict that damage to integral areas would sometimes produce phenomenologically striking deficits in visual consciousness. But a thorough survey of lesion data concerning the potential integral areas turns up no evidence—not even anecdotal—of the kinds

[41] An interesting question here concerns what effect the argument here has on another currently popular theory of consciousness, the global neuronal workspace theory (GNW) (Dehaene et al., 2006, 2014): In GNW, the dlPFC plays an important role, insofar as it is part of a network of areas (one also including the inferior parietal and inferotemporal cortices) that are said to be responsible for consciousness. This raises the question as to whether GNW has to predict that dlPFC impairment will cause conscious deficits. While space limits prevent us from examining this question closely, there is one consideration discussed above that makes it seem like GNW would have to predict dramatic deficits to occur, this being the idea that damaging any single node in a network is often sufficient for causing functional deficits (see Sect. 2).

of conscious deficit to be expected in cases of integral area impairment. It appears, furthermore, that each of the ways that the HOT theorist might try to explain the failure to find the deficits is deeply problematic. And so it seems that the data presented in this article—the legion of lesions, if you will—count strongly against HOT theory.[42]

## References

Adam, R., Johnston, K., & Everling, S. (2019). Recovery of contralesional saccade choice and reaction time deficits after a unilateral endothelin-1-induced lesion in the macaque caudal prefrontal cortex. *Journal of Neurophysiology, 122*(2), 672–690.

Adam, R., Johnston, K., Menon, R. S., & Everling, S. (2020). Functional reorganization during the recovery of contralesional target selection deficits after prefrontal cortex lesions in macaque monkeys. *NeuroImage, 207*, 116339.

Ainsworth, M., Browncross, H., Mitchell, D. J., Mitchell, A. S., Passingham, R. E., Buckley, M. J., Duncan, J., & Bell, A. H. (2018). Functional reorganisation and recovery following cortical lesions: A preliminary study in macaque monkeys. *Neuropsychologia, 119*, 382–391.

Andersen, R. A., Asanuma, C., & Cowan, W. M. (1985). Callosal and prefrontal associational projecting cell populations in area 7A of the macaque monkey: A study using retrogradely transported fluorescent dyes. *The Journal of Comparative Neurology*. https://doi.org/10.1002/cne.902320403

Andrewes, D. (2001). Neuropsychology: From theory to practice. Psychology Press.

Baier, B., Geber, C., Müller-Forell, W., Müller, N., Dieterich, M., & Karnath, H.-O. (2015). Anosognosia for obvious visual field defects in stroke patients. *Brain Structure and Function, 220*(3), 1855–1860. https://doi.org/10.1007/s00429-014-0753-5

Bamiou, D.-E., Musiek, F. E., & Luxon, L. M. (2003). The insula (Island of Reil) and its role in auditory processing: Literature review. *Brain Research Reviews, 42*(2), 143–154.

Barbas, H., & Mesulam, M.-M. (1981). Organization of afferent input to subdivisions of area 8 in the rhesus monkey. *The Journal of Comparative Neurology*. https://doi.org/10.1002/cne.902000309

Barcelo, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience, 3*(4), 399.

Bartolomeo, P. (2021). From competition to cooperation: Visual neglect across the hemispheres. *Revue Neurologique, 177*(9), 1104–1111.

Bartolomeo, P., Thiebaut de Schotten, M., & Doricchi, F. (2007). Left unilateral neglect as a disconnection syndrome. *Cerebral Cortex, 17*(11), 2479–2490.

Beeckmans, J. (2007). Can higher-order representation theories pass scientific muster? *Journal of Consciousness Studies, 14*(9–10), 90–111.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind'and cognition. *Brain, 127*(4), 914–928.

Bisiach, E., Vallar, G., Perani, D., Papagno, C., & Berti, A. (1986). Unawareness of disease following lesions of the right hemisphere: Anosognosia for hemiplegia and anosognosia for hemianopia. *Neuropsychologia, 24*(4), 471–482. https://doi.org/10.1016/0028-3932(86)90092-8

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

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis, 71*(3), 419–431. https://doi.org/10.1093/analys/anr037

Block, N. (2009). Comparing the major theories of consciousness.

[42] In the case of HO theories that are not forms of HOT theory (Lycan 1996; Lau 2019; Cleeremans 2011) (see fn. 7), the question of whether what was argued here disconfirms them as well largely turns on whether the kind of HO state that these theories take to be constitutive of consciousness can be produced by areas outside of the potential integral areas considered in this article. We lack space for entering into this issue here.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *Journal of Neuroscience, 37*(40), 9603–9613.

Bonini, F., Lambert, I., Wendling, F., McGonigal, A., & Bartolomei, F. (2016). Altered synchrony and loss of consciousness during frontal lobe seizures. *Clinical Neurophysiology, 127*(2), 1170–1175.

Brickner, R. M. (1936). *The intellectual functions of the frontal lobes*. Macmillan.

Brickner, R. M. (1952). Brain of patient A. after bilateral frontal lobectomy status of frontal-lobe problem. *AMA Archives of Neurology & Psychiatry, 68*(3), 293–313.

Brogaard, B. (2011a). Are there unconscious perceptual processes? *Consciousness and Cognition, 20*(2), 449–463.

Brogaard, B. (2011b). Conscious vision for action versus unconscious vision for action? *Cognitive Science, 35*(6), 1076–1104.

Brown, R. (2015). The HOROR theory of phenomenal consciousness. *Philosophical Studies, 172*(7), 1783–1794.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences., 23*, 754.

Cairns, H., Oldfield, R. C., Pennybacker, J. B., & Whitteridge, D. (1941). Akinetic mutism with an epidermoid cyst of the 3rd ventricle. *Brain, 64*(4), 273–290.

Carruthers, P. (1999). Sympathy and subjectivity. *Australasian Journal of Philosophy, 77*(4), 465–482.

Carruthers, P. (1992). The animals issue: Moral theory in practice. Cambridge University Press.

Carruthers, P. (2000). Phenomenal consciousness: A naturalistic theory. Cambridge University Press.

Cazzoli, D., Schumacher, R., Baas, U., Müri, R. M., Wiest, R., Bohlhalter, S., Hess, C. W., & Nyffeler, T. (2012). Bilateral neglect after bihemispheric strokes. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *48*(4).

Chiang, T.-C., Lu, R.-B., Hsieh, S., Chang, Y.-H., & Yang, Y.-K. (2014). Stimulation in the dorsolateral prefrontal cortex changes subjective evaluation of percepts. *PLoS ONE, 9*(9), e106943.

Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2011.00086

Cullum, C. M. (1998). Neuropsychological Assessment of Adults. In *Comprehensive Clinical Psychology*. Elsevier. https://doi.org/10.1016/B0080-4270(73)00227-3

Damasio, A. (1994). *Descartes' error*. Random House.

Damasio, A., & van Hoesen, G. (1983). Focal lesions of the limbic frontal lobe. In *Neuropsychology of human emotion* (pp. 85–110). Guilford Press.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

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., Charles, L., King, J. R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology, 25*, 76–84. https://doi.org/10.1016/j.conb.2013.12.005

de Graaf, T. A., de Jong, M. C., Goebel, R., van Ee, R., & Sack, A. T. (2011). On the functional relevance of frontal cortex for passive and voluntarily controlled bistable vision. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhr015

del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain, 132*(9), 2531–2540.

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain, 118*(1), 279–306.

Dretske, F. I. (1995). *Naturalizing the mind*. MIT.

Farne, A., Buxbaum, L. J., Ferraro, M., Frassinetti, F., Whyte, J., Veramonti, T., Angeli, V., Coslett, H. B., & Ladavas, E. (2004). Patterns of spontaneous recovery of neglect and associated disorders in acute right brain-damaged patients. *Journal of Neurology, Neurosurgery & Psychiatry, 75*(10), 1401–1410.

Feuchtwanger, E. (1923). *Die funktionen des Stirnhirns ihre pathologie und psychologie* (Vol. 38). Springer-Verlag.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain, 137*(10), 2811–2822.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology, 37*, 66–74.

Gainotti, G., Almonti, S., di Betta, A. M., & Silveri, M. C. (1998). Retrograde amnesia in a patient with retrosplenial tumour. *Neurocase, 4*(6), 519–526.

Gennaro, R. (1993). Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers, 22*(1), 51–69.

Gennaro, R. (1996). *Consciousness and self-consciousness: A defense of the higher-order thought theory of consciousness* (Vol. 6). John Benjamins Publishing.

Gennaro, R. (2012). The consciousness paradox. *Consciousness, Concepts, and Higher-Order Thoughts, Cambridge*.

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia, 33*(5), 623–642.

Grafman, J. (2000). Conceptualizing functional neuroplasticity. *Journal of Communication Disorders, 33*(4), 345–356.

Grafman, J., & Litvan, I. (1999). Evidence for four forms of neuroplasticity. In *Neuronal plasticity: Building a bridge from the laboratory to the clinic* (pp. 131–139). Springer.

Grafman, J., Zahn, R., & Wassermann, E. (2010). Brain damage: Functional reorganization. In *Encyclopedia of Neuroscience* (pp. 327–331). Elsevier Ltd.

Heath, R. G., Carpenter, M. B., Mettler, F. A., & Kline, N. S. (1949). Visual apparatus: Visual fields and acuity, color vision, autokinesis. *Selective Partial Ablation of the Frontal Cortex, a Correlative Study of Its Effects on Human Psychotic Subjects*, 489–491.

Hebben, N., & Milberg, W. (2009). *Essentials of neuropsychological assessment* (Vol. 70). Wiley.

Heider, B. (2000). Visual form agnosia: Neural mechanisms and anatomical foundations. *Neurocase, 6*(1), 1–12.

Hill, C. (2011). How to study introspection. *Journal of Consciousness Studies, 18*(1), 21.

Husain, M., & Kennard, C. (1996). Visual neglect associated with frontal lobe infarction. *Journal of Neurology, 243*(9), 652–657.

Jacob, P., & de Vignemont, F. (2010). Spatial coordinates and phenomenology in the two-visual systems model. *Perception, Action, and Consciousness: Sensorimotor Dynamics and Two-Visual Systems*, 125–144.

Jerde, T. A., & Curtis, C. E. (2013). Maps of space in human frontoparietal cortex. *Journal of Physiology-Paris, 107*(6), 510–516.

Karnath, H.-O. (2015). Spatial attention systems in spatial neglect. *Neuropsychologia, 75*, 61–73.

Kasahata, N., Kawamura, M., Shiota, J., Araki, S., & Sugita, K. (1994). A case of verbal amnesia due to left retrosplenial lesion. *Japanese Journal of Stroke, 16*, 290–295.

Katai, S., Maruyama, T., Hashimoto, T., & Yanagisawa, N. (1992). A case of cerebral infarction presenting as retrosplenial amnesia. *Rinsho Shinkeigaku Clinical Neurology, 32*(11), 1281–1287.

Kozuch, B. (2014). Prefrontal lesion evidence against higher-order theories of consciousness. *Philosophical Studies, 167*(3), 721–746.

Kozuch, B. (2021). Underwhelming force: Evaluating the neuropsychological evidence for higher-order theories of consciousness. *Mind & Language, 37*(5), 790–813. https://doi.org/10.1111/mila.12363

Kozuch, B. (2018). Gorillas in the missed (but not the unseen): Reevaluating the evidence for attention being necessary for consciousness. *Mind & Language, 34*(3), 299–316. https://doi.org/10.1111/mila.12216

Kozuch, B. (2022). Conscious vision guides motor action—rarely. *Philosophical Psychology*, 1–34.

Krakauer, J. W. (2006). Motor learning: Its relevance to stroke recovery and neurorehabilitation. *Current Opinion in Neurology, 19*(1), 84–90. https://doi.org/10.1097/01.wco.0000200544.29915.cc

Kriegel, U. (2007). A cross-order integration hypothesis for the neural correlate of consciousness. *Consciousness and Cognition, 16*(4), 897–912.

Lau, H., & Brown, R. (2019). The Emperor's New Phenomenology? The Empirical Case for Conscious Experiences without First-Order Representations. In A. Pautz & D. Stoljar (Eds.), *Blockheads! Essays on Ned Block's philosophy of mind and consciousness* (pp. 199–213). MIT.

Lau, H., & Passingham, R. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences, 103*(49), 18763–18768.

Lau, H. (2010). Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(1), 1–7.

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

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

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences, 114*(10), E2016–E2025.

Lee, T. G., & D'Esposito, M. (2012). The dynamic nature of top-down signals originating from prefrontal cortex: A combined fMRI–TMS study. *Journal of Neuroscience, 32*(44), 15458–15466.

Lew, S., & Lau, H. (2017). Crucial role of the prefrontal cortex in conscious perception. In *Executive functions in health and disease* (pp. 129–141). Elsevier.

Lycan, W. (1996). *Consciousness and experience*. Cambridge: MIT Press.

Mackey, W. E., Devinsky, O., Doyle, W. K., Meager, M. R., & Curtis, C. E. (2016). Human dorsolateral prefrontal cortex is not necessary for spatial working memory. *Journal of Neuroscience, 36*(10), 2847–2856.

Markowitsch, H. J., & Kessler, J. (2000). Massive impairment in executive functions with partial preservation of other cognitive functions: the case of a young patient with severe degeneration of the prefrontal cortex. In *Executive Control and the Frontal Lobe: Current Issues* (pp. 94–102). Springer.

McCarthy, R., & Warrington, E. (1986). Visual associative agnosia: A clinico-anatomical study of a single case. *Journal of Neurology, Neurosurgery & Psychiatry, 49*(11), 1233–1240.

Metcalfe, J. & Schwartz, B. L. (2015). The ghost in the machine: Self-reflective consciousness and the neuroscience of metacognition. In J. Dunlosky & S. Tauber (Eds.), *The Oxford handbook of metacognition* (pp. 407–424). Oxford: Oxford University Press.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

Miyamoto, K., Osada, T., Setsuie, R., Takeda, M., Tamura, K., Adachi, Y., & Miyashita, Y. (2017). Causal neural network of metamemory for retrospection in primates. *Science, 355*(6321), 188–193.

Morales, J., & Lau, H. (2020). The Neural Correlates of Consciousness. In *Oxford Handbook of the Philosophy of Consciousness*. OUP.

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences, 8*(3), 102–107.

Nudo, R. J. (2011). Neural bases of recovery after brain injury. *Journal of Communication Disorders, 44*(5), 515–520. https://doi.org/10.1016/j.jcomdis.2011.04.004

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *Journal of Neuroscience, 37*(40), 9593–9602.

Owen, A. M. (1997). The functional organization of working memory processes within human lateral frontal cortex: The contribution of functional neuroimaging. *European Journal of Neuroscience, 9*(7), 1329–1339.

Penfield, W., & Evans, J. P. (1935). The frontal lobe in man: a clinical study of maximum removals. *Brain: A Journal of Neurology, 58*, 115–133.

Perlstein, W. M., Cole, M. A., Demery, J. A., Seignourel, P. J., Dixit, N. K., Larson, M. J., & Briggs, R. W. (2004). Parametric manipulation of working memory load in traumatic brain injury: Behavioral and neural correlates. *Journal of the International Neuropsychological Society, 10*(5), 724–741.

Philiastides, M. G., Auksztulewicz, R., Heekeren, H. R., & Blankenburg, F. (2011). Causal role of dorsolateral prefrontal cortex in human perceptual decision making. *Current Biology, 21*(11), 980–983.

Philippi, C. L., Bruss, J., Boes, A. D., Albazron, F. M., Deifelt Streese, C., Ciaramelli, E., Rudrauf, D., & Tranel, D. (2021). Lesion network mapping demonstrates that mind-wandering is associated with the default mode network. *Journal of Neuroscience Research, 99*(1), 361–373.

Pierrot-Deseilligny, C. H., Gray, F., & Brunet, P. (1986). Infarcts of both inferior parietal lobules with impairment of visually guided eye movements, peripheral visual inattention and optic ataxia. *Brain, 109*(1), 81–97.

Pollen, D. A. (2007). Fundamental requirements for primary visual perception. *Cerebral Cortex, 18*(9), 1991–1998.

Pollen, D. A. (1995). Cortical areas in visual awareness.

Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memory. *Fleeting Memories: Cognition of Brief Visual Stimuli*, 13–46.

Prinz, J. (2004). The fractionation of introspection. *Journal of Consciousness Studies, 11*(7–8), 40–57.

Ptak, R., & Schnider, A. (2011). The attention network of the human brain: Relating structural damage associated with spatial neglect to functional imaging correlates of spatial attention. *Neuropsychologia, 49*(11), 3063–3070.

Ptak, R., Schnider, A., & Fellrath, J. (2017). The dorsal frontoparietal network: A core system for emulated action. *Trends in Cognitive Sciences, 21*(8), 589–599.

Quraishi, I. H., Benjamin, C. F., Spencer, D. D., Blumenfeld, H., & Alkawadri, R. (2017). Impairment of consciousness induced by bilateral electrical stimulation of the frontal convexity. *Epilepsy & Behavior Case Reports, 8*, 117–122.

Raccah, O., Block, N., & Fox, K. C. R. (2021). Does the Prefrontal Cortex Play an Essential Role in Consciousness? Insights from Intracranial Electrical Stimulation of the Human Brain. *The Journal of Neuroscience, 41*(10), 2076–2087. https://doi.org/10.1523/JNEUROSCI.1141-20.2020

Ramsey, L. E., Siegel, J. S., Baldassarre, A., Metcalf, N., Zinn, K., Shulman, G. L., & Corbetta, M. (2016). Normalization of network connectivity in hemispatial neglect recovery. *Annals of Neurology, 80*(1), 127–141.

Reitan, R., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery*. Neuropsychology Press.

Rengachary, J., He, B. J., Shulman, G., & Corbetta, M. (2011). A behavioral analysis of spatial neglect and its recovery after stroke. *Frontiers in Human Neuroscience, 5*, 29.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497*(7451), 585–590.

Rizzolatti, G., Matelli, M., & Pavesi, G. (1983). Deficits in attention and movement following the removal of postarcuate (area 6) and prearcuate (area 8) cortex in macaque monkeys. *Brain, 106*(3), 655–673.

Ro, T., Cohen, A., Ivry, R. B., & Rafal, R. D. (1998). Response channel activation and the temporoparietal junction. *Brain and Cognition, 37*(3), 461–476.

Rosenthal, D. (2011). Exaggerated reports: Reply to Block. *Analysis, 71*(3), 431–437.

Rosenthal, D. (2002). Explaining consciousness. *Philosophy of Mind: Classical and Contemporary Readings, 46*, 406–421.

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

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). Theory of mind'impairments and their relationship to executive functioning following frontal lobe excisions. *Brain, 124*(3), 600–616.

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience, 7*(5), 499.

Sanchez-Carrion, R., Fernandez-Espejo, D., Junque, C., Falcon, C., Bargallo, N., Roig, T., Bernabeu, M., Tormos, J. M., & Vendrell, P. (2008). A longitudinal fMRI study of working memory in severe TBI patients with diffuse axonal injury. *NeuroImage, 43*(3), 421–429.

Schiller, P. H., & Chou, I. (1998). The effects of frontal eye field and dorsomedial frontal cortex lesions on visually guided eye movements. *Nature Neuroscience, 1*(3), 248–253.

Schiller, P. H., & Chou, I. (2000). The effects of anterior arcuate and dorsomedial frontal cortex lesions on visually guided eye movements: 2. *Paired and Multiple Targets. Vision Research, 40*(10–12), 1627–1638.

Schwitzgebel, E. (2012). Introspection, what? Introspection and Consciousness, 29–48.

Sebastián, M. Á. (2014). Not a HOT dream. In *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (pp. 415–432). Springer.

Serino, A., Ciaramelli, E., di Santantonio, A., Malagù, S., Servadei, F., & Làdavas, E. (2006). Central executive system impairment in traumatic brain injury. *Brain Injury, 20*(1), 23–32.

Siewert, C. (1998). *The significance of consciousness*. Princeton University Press.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*(5), 640–656.

Szczepanski, S. M., & Knight, R. T. (2014). Insights into human behavior from lesions to the prefrontal cortex. *Neuron, 83*(5), 1002–1018.

Takahashi, N., Kawamura, M., Shiota, J., Kasahata, N., & Hirayama, K. (1997). Pure topographic disorientation due to right retrosplenial lesion. *Neurology, 49*(2), 464–469.

Thompson-Schill, S. L., Jonides, J., Marshuetz, C., Smith, E. E., D'Esposito, M., Kan, I. P., Knight, R. T., & Swick, D. (2002). Effects of frontal lobe damage on interference effects in working memory. *Cognitive, Affective, & Behavioral Neuroscience, 2*(2), 109–120.

Toba, M. N., Migliaccio, R., Batrancourt, B., Bourlon, C., Duret, C., Pradat-Diehl, P., Dubois, B., & Bartolomeo, P. (2018). Common brain networks for distinct deficits in visual neglect. A combined structural and tractography MRI approach. *Neuropsychologia, 115*, 167–178.

Turatto, M., Sandrini, M., & Miniussi, C. (2004). The role of the right dorsolateral prefrontal cortex in visual change awareness. *NeuroReport, 15*(16), 2549–2552.

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances, 2*, 2398212818810591.

Valenstein, E., Bowers, D., Verfaellie, M., Heilman, K. M., Day, A., & Watson, R. T. (1987). Retrosplenial amnesia. *Brain, 110*(6), 1631–1646.

Voytek, B., Davis, M., Yago, E., Barceló, F., Vogel, E. K., & Knight, R. T. (2010). Dynamic neuroplasticity after human prefrontal cortex damage. *Neuron, 68*(3), 401–408.

Voytek, B., & Knight, R. T. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences, 107*(42), 18167–18172.

Vuilleumier, P. (2004). Anosognosia: The neurology of beliefs and uncertainties. *Cortex, 40*(1), 9–17.

Weisberg, J. (2011). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis, 71*(3), 438–443. https://doi.org/10.1093/analys/anr040

Weiskrantz, L. (1986). Blindsight: A case study and implications.

Yago, E., Duarte, A., Wong, T., Barceló, F., & Knight, R. T. (2004). Temporal kinetics of prefrontal modulation of the extrastriate cortex during visual attention. *Cognitive, Affective, & Behavioral Neuroscience, 4*(4), 609–617.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain, 113*(6), 1721–1777.

Zihl, J., von Cramon, D., & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain, 106*(2), 313–340.

Zimmer, H. D. (2008). Visual and spatial working memory: From boxes to networks. *Neuroscience & Biobehavioral Reviews, 32*(8), 1373–1395.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.