Nathaniel Greely

Higher-Order Theories of Consciousness are Empirically False

Abstract: *Higher-order theories of consciousness come in many varieties, but all adopt the 'transitivity principle' as a central, explanatory premise. The transitivity principle states that a mental state of a subject is conscious if and only if the subject is aware of it. This higher-order awareness is realized in different ways in different forms of higher-order theory. I argue that empirical studies of meta-cognition have falsified the transitivity principle by showing that there can be awareness of a mental state without that state's becoming conscious. I present two such studies in detail and argue that the measures they employ cannot be interpreted in a way that would make the results compatible with higher-order theory. Since all versions of the theory rely on the transitivity principle, this entails that all forms of higher-order theory are false.*

1. Introduction

The higher-order theory of consciousness (HO) just won't die. The venerable theory, which dates at least to Kant (1781/1996), rose to prominence in the late twentieth century through the work of Armstrong (1968), Rosenthal (1986; 2005), Lycan (1996), and others. Although its demise has been proclaimed repeatedly (e.g. Block, 2011; Sauret and Lycan, 2014), HO continues to gain adherents and new variants continue to crop up (e.g. Brown, 2015; Coleman, 2015; Gennaro, 2012). HO has recently gained a toehold in cognitive science (Brown, Lau and LeDoux, unpublished; Cleeremans, 2011; LeDoux

Correspondence:
University of California San Diego, Department of Philosophy, 9500 Gilman Dr. #0119, La Jolla, CA, 92092, USA. *Email: ngreely@ucsd.edu*

Journal of Consciousness Studies, **27**, No. 11–12, 2020, pp. 30–54

and Brown, 2017), invigorating some of HO's proponents, who now claim empirical support for the theory (Lau and Rosenthal, 2011). This development offers an opportunity for HO's opponents as well. In this article I will argue that careful analysis of the empirical evidence shows that HO is false.

One reason for HO's survival in the face of sustained critique is that the theory is hydra-headed. HO comprises a family of theories sharing the core claim that a mental state is conscious if and only if it is the intentional object of another mental state.[1] Call these consciousness-making states *higher-order representations* (HORs). Various members of the HO family invoke different types of HORs, most prominently *higher-order thought theory* (HOT) and *higher-order perception theory* (HOP). This variety has made HO an unwieldy beast for those who wish to slay it.

However, there is a claim to which all versions adhere, as it provides HO's explanatory core — the *transitivity principle* (TP) (Rosenthal, 1997). As formulated by Lycan, it states that 'a conscious state is a mental state whose subject is aware of being in it' (2001, p. 3). According to TP, awareness of a mental state is both necessary and sufficient for that state's being conscious. TP is more fundamental to HO than any specific claim about the nature of the representation that provides this awareness. Thus, if TP can be refuted, HO will topple.

Meanwhile, the empirical study of higher-order representation, or *metacognition*, has grown by leaps and bounds as new statistical methods grounded in signal detection theory have allowed researchers to more effectively distinguish first- and higher-order representations behaviourally (Galvin *et al.*, 2003; Maniscalco and Lau, 2012). This has made it possible to test certain predictions of HO. Recent studies, I will argue, prove one prediction false — awareness of a mental state is not sufficient for that state's being conscious (e.g. Charles *et al.*, 2013; Jachs *et al.*, 2015).

In Section 2 I will present the basic tenets of HO, emphasizing the way in which HO depends on TP for its motivation and explanatory power. In Section 3 I will present two studies that falsify TP's prediction that awareness of a mental state is sufficient for consciousness. In Section 4 I will defend my interpretation of these studies, arguing that the measures of consciousness and metacognitive awareness

[1] Some HO theorists attribute consciousness to subjects rather than mental states (e.g. Berger, 2014; Brown, 2015; Rosenthal, 2011). More on this later.

employed pick out the same phenomena invoked by TP. In Section 5 I will argue that there are no defects in the relation between the higher-order mental state and its object in these studies, thus blocking a potential line of defence for proponents of HO. In Section 6 I will consider some objections Rosenthal (2011; 2012; 2019) has lodged against the identification of higher-order awareness with metacognition and argue that they cannot save HO.

2. HO and How to Refute it

> Explaining what it is in virtue of which conscious states differ from mental states that aren't conscious is the principal goal of a theory of consciousness. And it's fairly straightforward to get a start on that question. When a mental state is conscious, the individual that's in that state is conscious of it; when a mental state fails to be conscious, that individual is in no way whatever conscious of that state. (Rosenthal, 2005, p. 3)

Here Rosenthal presents the foundation of his explanation of consciousness, which he calls the *transitivity principle* (TP). The principle is also adopted by Lycan as the first premise in his 'simple' argument for HO. Lycan substitutes an identity claim for the biconditional and 'aware of' for one instance of 'conscious', but the similarity is obvious:

> (1) A conscious state is a mental state whose subject is aware of being in it. (Lycan, 2001, p. 3)

Gennaro formulates TP as:

> A conscious state is a state whose subject is, in some way, aware of being in it. (2012, p. 28)

TP provides the explanatory core of HO. The move is subtle, striking some as a conceptual truth. The necessity claim is plausible. If one is conscious of something, it seems one must be aware of it. But much hinges on what we mean by 'aware'. HO theorists define 'aware' as an intentional relation, concluding that higher-order representation of a mental state is both necessary *and* sufficient for that state's being conscious (e.g. Rosenthal, 1986, p. 335; Lycan, 2001, p. 4). Thus, consciousness *simpliciter* is explained in terms of the specific phenomenon of *transitive* consciousness. All versions of HO rely on this essential explanatory move (Rosenthal and Weisberg, 2008, para. 8).

There is disagreement among HO theorists over how TP should be formulated, and the conclusion that results. For some, TP explains *state consciousness* (Gennaro, 2012; Wilberg, 2010). Others take the explanandum to be *subject consciousness* (Brown, 2015; Rosenthal, 2011). The different interpretations of TP produce two versions of HO:

> **State-HO.** A mental state is conscious if and only if it is the intentional object of a higher-order mental state (of the right type and in the right way).
> **Subject-HO.** A subject is conscious if and only if she is represented as being in a first-order mental state by a higher-order mental state (of the right type and in the right way).

The parenthetical hedging is worked out in the details of the various versions of HO. *Higher-order perception theory* (HOP) takes the relevant HOR to be quasi-perceptual (Armstrong, 1968; Lycan, 1996). *Higher-order thought theory* (HOT) holds that the HOR must be a *thought* and that the way it is related to the first-order state must be *non-inferential* (Rosenthal, 2005). The non-inferentiality condition rules out situations in which HOTs obtain via conscious inference — your therapist might convince you that you are angry, and she may be right, but this does not entail that you will consciously *feel* angry. Other HO theorists take the first- and higher-order content to be parts of a single, complex mental state (Gennaro, 2012). Finer distinctions have been made within types of HO, including dispositional HOT (Carruthers, 1996; 2004), quotational HOT (Coleman, 2015), HOROR (Brown, 2015), and so on, and each may have different constraints on the way in which the consciousness-making relation obtains. But all varieties of HO rely on TP as the fundamental explanatory move.

HO, in virtue of its adherence to TP, makes some empirical predictions. It predicts that consciousness will not obtain in the absence of the requisite HOR. It also predicts that when this HOR does obtain, its object (be it a mental state or a subject) will necessarily be conscious. In Section 3 I will present studies that falsify the latter prediction, showing that consciousness can fail to obtain in this circumstance. These studies are not couched in philosophical terms and it will require a bit of interpretation to show that they falsify HO. A little preparation now will ease that task.

First, the fundamentality of TP limits HO theorists' options in interpreting these studies. Empirical studies of metacognition do not distinguish as finely among types of higher-order representations as

do philosophers. That is, metacognitive measures are frequently agnostic as to whether the HOR purportedly assessed is a thought, a perception, or something else. So, if faced with a study that claims that HORs obtain in the absence of consciousness, couldn't a HOP theorist just respond that those HORs are probably HOTs, in which case HOP is left untouched? No, because the consciousness-making property, according to TP, is the awareness itself. HO theorists favour their types of HORs because they think that they are particularly suited to provide this awareness (Rosenthal and Weisberg, 2008, para. 8). For example, Lycan (2004) argues that thought *presupposes* awareness, and that HOPs are the source of that awareness. For him to concede that HOTs occur in the absence of consciousness would be to concede the same for HOPs. Likewise, Rosenthal (2005; Rosenthal and Weinberg, 2008), rather than argue that HOPs provide the wrong kind of awareness, argues that the purported mechanism is incapable of producing awareness at all, since mental states lack perceptible qualities. So it is not necessary to nail down the type of HOR measured in the studies I will present. I can bypass that whole kerfuffle and instead attempt to determine whether these measures assess awareness of the subject's first-order states *á la* TP. In Section 4 I will argue that they do.

But the parenthetical in HO has two conjuncts. The HO theorist might concede the first and instead attempt to dispute these studies on the grounds that the higher-order states involved are not related to their objects in the right *way*. For example, a HOT theorist might concede that the studies do show that a HOT can exist in the absence of consciousness but argue that the metacognitive relation to its intentional object is defective in some way, for example by involving a conscious inference. Here the response is not so simple. To respond to each potential condition cited by each type of HO would take too much space. What I will attempt to do instead is take down the biggest guy in the room. I take HOT to be the strongest extant HO theory, since the only living proponent of HOP has given it up (Sauret and Lycan, 2014), and I take Rosenthal's version to be at the centre of the current interdisciplinary interest in HO (see, for example, Brown, Lau and LeDoux, unpublished, p. 4). In Section 5 I will consider the ways a HOT theorist might argue that the HORs in the studies I'll present don't relate to their objects in the right way and I will argue that these responses fail. To refute HOT empirically is a worthwhile project in its own right, but the responses I consider on behalf of HOT are in many cases similar to ones that would be offered by other types of HO

theorists, thus sketching a blueprint for the refutation of remaining versions of HO. Finally, in Section 6 I will address some of Rosenthal's comments about the relation between metacognition as studied in cognitive science and the higher-order awareness invoked by TP.

3. Evidence Against HO

All forms of HO are committed to the claim that a subject's higher-order awareness of first-order mental content entails consciousness of that content. This claim is more fundamental than whether the higher-order state is a thought, a perception, or anything else. Empirical research on 'metacognition', psychologists' term for higher-order mental processes, has shown that such awareness can obtain in the absence of said consciousness. I will present two such studies in detail (Charles *et al.*, 2013; Jachs *et al.*, 2015).

Charles *et al*. (2013) use a standard psychophysical paradigm to measure three variables — (a) first-order (or 'type-1') perception of a stimulus, (b) conscious awareness of that stimulus, and (c) meta-cognitive (or 'type-2') awareness. They also use electro- and magneto-encephalography (EEG/MEG) to measure error detection, a metacog-nitive ability, but for simplicity we can focus on the behavioural aspect of the study. The authors use a masking paradigm to create conscious and unconscious conditions. A number is displayed briefly on a screen. After the stimulus disappears, there is a variable delay, known as *stimulus onset asynchrony* (SOA), before the presentation of a visual 'mask'. The mask is a shape which appears in the area previously occupied by the stimulus, disrupting the retention of the stimulus in short-term memory (see Figure 1). Manipulation of SOAs creates conscious conditions at longer SOAs and unconscious con-ditions at shorter SOAs. After the presentation of the stimulus, sub-jects are asked, in a forced-choice task, whether the number displayed was greater or less than 5. The forced choice allows for the possibility that subjects may succeed at the type-1 task even when they deny any conscious experience of the stimulus. It is well-established that sub-jects can perform significantly above chance in such conditions. Con-scious awareness is then measured by asking subjects whether they

consciously saw the stimulus,[2] and metacognition is measured by asking subjects whether their type-1 response was correct or incorrect.

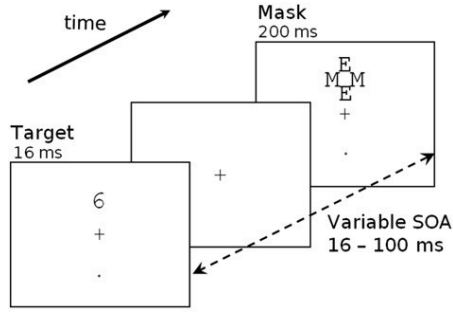

Figure 1. Stimulus and mask from Charles *et al.* (2013, p. 82).

Reported consciousness increased with SOA, indicating that subjects were better able to consciously see the stimulus when it was not immediately followed by the mask. Type-1 performance also increased as SOA increased, as did metacognitive accuracy. Statistical analysis also showed that metacognitive accuracy was greater in conscious trials than unconscious trials. These broad correlations between consciousness, task performance, and metacognitive accuracy are predictable and intuitive. In general, we are better at visual tasks when we are conscious of the stimulus, and we are better at judging our performance on those tasks when we are conscious of the stimulus.

But HO predicts more than a broad correlation between consciousness and metacognition. HO holds that higher-order awareness is sufficient for consciousness, and so when that awareness obtains, consciousness cannot fail to obtain as well. Charles *et al.* found not only that type-1 performance was significantly above chance in a significant portion of unconscious trials, but that metacognitive performance was also significantly above chance in these trials (*ibid.*, p. 86). This shows that metacognition and consciousness dissociate. One can

[2] These studies control for the possibility that differences in reports of consciousness across conditions reflect a shift in the subject's internal standard for a positive response, the response 'criterion', rather than a change in consciousness. In Jachs *et al.* (2015, below) there is no variable that would induce criterion shift, as conscious and unconscious trials are identical. Charles *et al.* manipulate SOA between trials and compensate by using statistical analysis to create a criterion-free measure of consciousness (2013, p. 84).

accurately detect a property of a stimulus, demonstrate metacognitive awareness of that accuracy, and still report no consciousness of the stimulus. If I can successfully argue that these three abilities equate to perception, higher-order awareness, and consciousness as TP conceives them, this straightforwardly refutes HO. But first more data.

Jachs *et al.* (2015) presented subjects with a Gabor grating (roughly, a series of parallel lines) slanted in either a leftward or rightward direction, followed by a mask. Rather than use SOAs to manipulate consciousness, the luminance of the grating was adjusted until consciousness ratings reached 50%. Subjects first reported whether they were conscious of the stimulus, followed by a report of the orientation of the grating (type-1 task), followed by a metacognitive confidence rating (type-2 task). Subjects displayed above-chance performance on both type-1 and type-2 tasks in trials where no consciousness of the stimulus was reported. Thus, once again, it seems that subjects can accurately detect a property of a stimulus and demonstrate metacognitive awareness of this without reporting any conscious experience of the stimulus. Other studies draw the same conclusion. In Section 4 I will briefly discuss one more (Kanai, Walsh and Tseng, 2010), but the methodology is more complex. I have chosen Charles *et al.* (2013) and Jachs *et al*. (2015) because they represent as straightforward a refutation of HO as one could ask for. The bulk of *my* work in this article will consist in interpreting the measures of perception, consciousness, and metacognition employed.

Thankfully, I needn't spend much time discussing measures of perception and consciousness. Type-1 tasks, like those in these studies, have long been accepted as measures of perception. HO theorists are motivated to agree because they need there to be such a thing as unconscious perception. It is only when perceptual states become the object of HORs that they become conscious, otherwise they function in the same way. Forced-choice detection tasks allow for this possibility. HO theorists will also agree that subjective report is our best available measure of consciousness (e.g. Rosenthal, 2019). HO theorists are motivated to accept subjective report over the most common alternatives — type-1 performance and type-2 performance. HO theorists reject type-1 performance as a measure of consciousness for the same reason that they accept it as a measure of perception — they need to allow for the possibility of unconscious perception. Researchers sometimes use type-2 performance as a proxy for consciousness, the thought being that consciousness of a stimulus causes confidence in one's type-1 response. But studies like those I have just

presented show that confidence and consciousness dissociate (as Rosenthal, 2019, is well aware).[3] HO theorists' only option, then, is to deny that confidence ratings measure higher-order awareness.

4. Confidence and HORs

The studies I have presented dissociate higher-order awareness (as indicated by confidence ratings) from consciousness (as indicated by reports of consciousness). But the HO theorist might question whether confidence ratings measure the same phenomenon invoked by TP. Here I will argue that they do. Call the states measured by confidence ratings *c-states*. I will argue that c-states are higher-order representations of first-order perceptual states. If so, they provide higher-order awareness of the sort HO theorists claim is necessary for conscious perception. And so long as there are no defeaters, such as conscious inferences, they should also be sufficient for conscious perception.

I'll begin by tackling the issue in three parts, first arguing that c-states are intentional states, second that they take other mental states as their objects, and third that those objects are perceptions. Then in Section 5 I will argue that there are no defeaters that might prevent consciousness from arising even though these three conditions are satisfied.

According to subject-HO, the intentional object of a HOR is not, strictly speaking, a mental state but a subject. This distinction matters a great deal in some contexts, but not in ours. State-HO and subject-HO are primarily at odds over the possibility of higher-order misrepresentation. Subject-HO preserves the intuition that we can have conscious experiences that misrepresent their first-order objects by taking the content of the HOR alone as sufficient for consciousness. And the *content* of a consciousness-making HOR will be the same for state-HO theorists as for subject-HO theorists. That the HOR represents the first-order state as a state of the subject is automatically implied in typical cases (Berger, 2014, p. 837; Rosenthal, 2011, p. 436).[4] So when there is no misrepresentation both versions of HO make the

[3] Charles *et al.* (2013) ask subjects to rate correctness, but they clearly consider this a type of metacognitive confidence rating (pp. 82, 90). All I need argue is that HORs obtain.

[4] It's possible to represent a mental state as not belonging to oneself, for example when it is attributed to someone else. But this is not a remotely plausible interpretation of what is going on in the studies I describe.

same prediction — that HORs with the right type of content are sufficient for consciousness of that content. I must show, then, that the measures of higher-order awareness used in these studies indicate the presence of HORs with the right sort of content. For convenience I may sometimes speak of a perceptual state as the intentional object of the HOR, and subject-HO theorists often do the same (e.g. Rosenthal, 2011). The relevant point is that the HOR represents a first-order perceptual state as obtaining — the subject comes along for the ride.

4.1. C-states are intentional states

Lau and Rosenthal classify c-states as HORs,[5] and thus concede at the very least that they are intentional states (2011, p. 370).[6] This view is held by philosophers whose accounts of metacognition are otherwise vastly different, from Proust (2015), who views c-states as non-conceptual states of a control system, to Carruthers (2011) who denies any introspective access to our own cognitive states. More generally, confidence ratings, as they are employed in psychology, epistemology, decision theory, and a wide variety of disciplines, are taken to measure belief, an intentional state (Schwitzgebel, 2015). I don't know of any relevant players who doubt that c-states are intentional states. There may be some eliminativists who would, but they aren't HO theorists, as HO requires mental representations.

Furthermore, if a wide variety of competent researchers assume that c-states are intentional states because they find it intuitive that they have a certain content, then these converging opinions constitute more than an argument from authority, they constitute data. This is because the data by which we test theories of intentionality tend to be intuitions about the content of a given state. For example, when informational semantics predicts that mental content is wildly disjunctive, we alter the theory, not our view of mental content. Thus, the consensus of philosophers and psychologists that c-states are metacognitive states, and are thus intentional, is as good an argument as we will get, since any formal theory of intentionality will have to accommodate it.

---

5 They use the term 'sensory metacognition' for the phenomenon, but the context is the discussion of Kanai, Walsh and Tseng (2010), who use confidence ratings to measure sensory metacognition.

6 More will be said on Rosenthal's view of confidence (esp. in his 2019). For now, I'm only claiming that he takes confidence to indicate the presence of an intentional state.

4.2. The intentional objects of c-states are mental states

Now I must argue that c-states provide awareness of our mental states — that they are HORs. First, confidence ratings are typically correct. That is, if the type-1 response is correct, confidence is typically high, and if the type-1 response is incorrect, confidence is typically low (Koriat, 2000). This is one important function of metacognition — correcting for first-order errors. C-states, then, must be receiving information from the perceptual state, the stimulus, or some other element in the causal chain. But which is the intentional object?

A c-state might take the stimulus itself as its object if the subject in a study of metacognition misunderstands the task. The distinction between reporting a property of the stimulus and reporting a property of one's judgment about the stimulus is not obvious, at first blush, to the average participant in these studies. Are subjects expressing confidence in their *judgment* that the stimulus was, say, a square, or are they expressing confidence that the *stimulus* was a square? This is a problem that researchers take seriously, and much work has gone into ruling this possibility out. Galvin *et al.* (2003) is the seminal work here. They criticize earlier studies that fail to disambiguate type-2 tasks from type-1 tasks, and their prescriptions for avoiding this flaw have been cited and employed nearly ubiquitously in metacognitive studies since. Techniques for preventing ambiguity include carefully wording the type-2 prompts and carefully instructing subjects on the distinction between the first- and second-order tasks. There is, of course, always opportunity to be sloppy in this regard, but recent studies, like those reported in Section 3, tend not have this flaw.

If a subject does understand the task, is she then competent to aim her confidence rating at the correct intentional object? This is the assumption of researchers, and it's a reasonable one, but there are semantic theories that might question it. An externalist will claim that on Twin Earth one can fail to think about water despite one's best intentions. But it often requires a very specific set-up to pry the verdicts of internalist and externalist semantics apart, and there is no reason to think that the relatively straightforward tasks presented in these studies constitute such a set-up. Likewise, theorists of perception may disagree over whether certain properties belong to the perception or the stimulus, and some might deny that one can accurately attribute, say, greenness to a perception. But in these studies subjects attribute accuracy or inaccuracy to the perception, something everyone agrees is a property of perceptions and not of objects.

Another state that intervenes between the presentation of the stimulus and the c-state is the type-1 response. In judging her type-1 response to be correct, could the subject be referring not to the perception but to her answer about the perception? The answer itself is a button-press, not a linguistic token, but perhaps the object of the c-state is the proposition thereby expressed.[7] This, however, implies the existence of a psychological token with the same content — the source of the response. If that mental state represents the perceptual state, then it is itself a HOR, sufficient for consciousness. If it represents the stimulus it is not. I will consider the latter possibility in Section 4.3.

Since the type-1 task is a forced choice, and subjects sometimes guess, could the source of the answer fail to be an intentional state and instead be a mere causal process? Like button-presses, causal states are not in themselves correct or incorrect, so a subject who understands the task is unlikely to be referring to such states.

And anyhow, I have used the general consensus that c-states target mental states to argue that they are intentional states, and the same reasoning can be used to establish the nature of their intentional objects. In this case the datum is the consensus of researchers that confidence ratings measure HORs.

4.3. C-states take perceptual states as their objects

The HOT theorist might accept that the intentional objects of c-states are mental states of the subject, but deny that, in the studies in question, they take the relevant perceptual state as their object. The syntax of the confidence report indicates that the object of the c-state is truth-apt, since confidence reports assess their accuracy, and this suggests that, if it's not a perception or a proposition, it might be a belief.

Perhaps the intentional object of the c-state is a perceptual belief caused by the perceptual state. This belief may then be the basis of the subject's response in the type-1 task. If this belief is the intentional object of the c-state, then HO entails that the belief will be conscious, but what does it entail about the perception? If the perceptual belief is caused by the perception, and more importantly takes the perception

7 Jachs *et al.* (2015) don't describe the format, but answers are numbers or single words.

as its intentional object, then there is a conscious-making HOR.[8] If so, HO still predicts that the percept will be conscious; the relevant HOR is simply pushed down into the type-1 task. But what if the perceptual belief takes the stimulus as its object? Then the HO theorist can deny that the c-state provides a sufficient condition for a conscious perceptual state. Instead we have an unconscious perceptual state, a conscious belief that is caused by the perceptual state but whose intentional object is the stimulus itself, and a HOR whose intentional object is the perceptual belief. Call this the 'belief gambit'.

The belief gambit is in tension with the standard interpretation of the type-1 task as a measure of perception, not of belief. There is reason for this. The time between the presentation of the stimulus and the type-1 response is very short, on the order of milliseconds, and the properties reported are typically low-level (e.g. orientation of a grating),[9] suggesting that higher cognitive processes have little time or reason to get involved. So when subjects are asked to evaluate their type-1 performance in the type-2 task, they are asked to evaluate their *perceptual* performance. If HO theorists want to reject this widely accepted interpretation, they should offer some reason.

Furthermore, in many of these studies there is no difference in experimental conditions between conscious and unconscious trials. In Jachs *et al*. (2015), for example, the stimulus is designed so as to produce a particular proportion of reports of consciousness vs. unconsciousness. It would be *ad hoc* to claim that perceptual beliefs arise only in unconscious trials. Instead, the HO theorist should claim that such perceptual beliefs obtain in *all* trials, forming the basis of confidence ratings, and that higher-order awareness, a separate phenomenon, only obtains in conscious trials. But because the standard interpretation of type-1 tasks has no need to posit such ubiquitous perceptual beliefs, the belief gambit will still seem *ad hoc* unless HO theorists can say something about their role.

HO theorists would need to adopt a substantive theory of perception on which perceptual states alone are insufficient to accomplish type-1 tasks. That is, they must claim that perception alone is not enough evidence for the subject to determine that the stimulus is a square. There is an affinity here with sense-data theory (Russell, 1912/1997).

8 Remember, HOP shouldn't deny this claim. See Section 2.

9 The type-1 task in Charles *et al*. (2013) requires recognition of a number, which might arguably be a higher-level property. That's one reason I present multiple studies.

Sense-data theory posits an inference from the properties of the perceptual state to a belief about the objects that cause it. The nature and reliability of this inference have been subjects of some debate in the past (e.g. Chisholm, 1950), and few hold a sense-data theory of perception now. But HO theorists who might opt for the belief gambit will find themselves in a similar situation to the sense-data theorist. If perception is insufficient for, but related to, type-1 performance, then the belief posited to fill the gap should infer something about the stimulus from the perception. But to perform such an inference would require multiple beliefs, at least one of which takes the perception as its object, which for the HOT theorist would entail the consciousness of that perception. To thread this needle with an account that is not *ad hoc*, empirically adequate, and doesn't end up entailing consciousness may be impossible. And even if it's not, it's a heavy commitment to a theory of perception that HO theorists shouldn't want to be saddled with, particularly if they want to retain HO's appeal in the cognitive sciences.[10]

In this section I have argued that, in the studies I have presented, c-states — the states picked out by confidence ratings — are intentional states whose intentional objects are mental states, specifically perceptions. According to HO these are sufficient conditions for the perceptions to be conscious. But in the studies I have presented, the perceptions are not conscious. HO theorists have one more trick up their sleeves, however, and I turn to this in the next section.

5. No Defeaters

If HO theorists must concede that c-states are HORs of first-order perceptual states, there is still another way out. The HO theorist might claim that the nature of the relation between the c-state and its object is somehow defective. There are too many brands of HO for me to provide an exhaustive account of these potential strategies, so I've taken Rosenthal's version of HOT to be the strongest adversary. Rosenthal has lately been engaged in the empirical argument for HO, offering a collection of arguments that various metacognitive measures do not provide higher-order awareness of the consciousness-inducing kind. I think that many of the points I will make in what

[10] In forced-choice tasks subjects may not endorse their responses, so the relevant mental states may not be beliefs. Whatever the nature of the mental states, if they serve the same functional role, the same arguments apply.

follows will apply, *mutatis mutandis*, to other versions of HO. If a proponent of some subtype of HO thinks I have missed something, I'd love to hear the argument, but I can't pre-empt them all. In this section I will consider some general objections that a proponent of Rosenthal's version of HOT might make to the nature of the relation between c-states and their objects.[11] In Section 6 I will consider some objections by Rosenthal himself.

5.1. Conscious inference

The intentional relation between the perception and the HOT might be defective if it is the result of a conscious inference. This non-inferentiality condition is not designed for the sorts of mundane perceptual episodes that occur in the studies I have cited. By and large, HOT theorists want to say that these are conscious. But could a HOT theorist plausibly claim that, in prompting subjects to report confidence ratings, researchers also prompt a conscious inference that would defeat consciousness?

One might claim that subjects make an inference based on observations of their own behaviour. There are theorists who believe that metacognition involves just such inferences, but the view is only plausible because they take these inferences to be automatic and unconscious (e.g. Carruthers, 2011).

And to invoke conscious inference as a consciousness-defeater would imply a difference between conscious and unconscious conditions in these studies. In conscious conditions, the HOT theorist will want to claim that the HOT does not run afoul of the non-inferentiality condition, but in unconscious conditions there is a conscious inference somewhere. The trials are so similar between conscious and unconscious conditions, in some cases identical, and happen so quickly that it is hard to see where, or why, the inference would occur. Since the posited inference must be conscious, we could simply ask the subjects if they notice one. I've tried it on myself and I don't. So more empirical work could be done to study this question, but the prospects for this line of response don't look good for HOT.

[11] Again, this talk of intentional objects should be taken, in the case of subject-HO, as talk of the intentional content of the HOT.

5.2. Temporal relations

In the studies presented, the confidence rating is given after the stimulus has disappeared. A state-HOT theorist (though not a subject-HOT theorist) might object, then, that the relation is thereby defective because one of the relata, the first-order perceptual state, no longer obtains. Rosenthal (2012) has argued against identifying metacognitive states like tip-of-the-tongue experiences, judgments of learning, and feelings of knowing with HOTs on similar grounds, as these assess one's ability to access mental content at some point in the future. But in our case the objection only works if the HOT theorist can deny that a confidence rating is reliable evidence that the appropriate relation between the perception and the c-state obtained while the stimulus was present. In these studies, the temporal gap between stimulus and confidence rating is on the order of milliseconds, well within the range of working memory. Furthermore, confidence ratings tend to accurately assess first-order performance, suggesting a causal relation between the perception of the stimulus and the confidence rating. Whatever minimal temporal delay might obtain between cause and effect here will not be objectionable to the state-HOT theorist, as they typically describe the first-order state as causing the HOT, and thus embrace at least that degree of temporal lag.

5.3. The veridicality of the HOT

When subjects report their c-states through confidence ratings, these c-states are presumably conscious. Rosenthal holds that HOTs, by and large, are not conscious and only become conscious when they are the object of third-order thoughts. So any time a subject makes a confidence report based on a conscious second-order mental state, there is presumably a third-order HOT involved. But on Rosenthal's theory, there needn't be agreement in content between the higher-order state and the state to which it refers (2005, p. 217). If a first-order perceptual state is of a square, but this state is targeted by a higher-order thought with the content 'I am perceiving a diamond', the resulting conscious experience will be of a diamond. Other HOT theorists, like Gennaro (2012), find this possibility deeply problematic and tailor their theories to avoid this possibility. But Rosenthal could claim that, in any given trial, it is possible that the reported confidence misrepresents the content of the c-state, instead reporting the content of its third-order consciousness-maker. Given enough of this sort of misrepresentation, the studies I have cited might be invalid. The obvious

problem with this strategy, and part of the reason that HOT theorists like Gennaro deny Rosenthal's account, is that if misrepresentation of this sort were common, our conscious lives would break radically with reality, causing cognitive and behavioural chaos. So the plausibility of Rosenthal's theory relies on this sort of misrepresentation being minimal, but to use it as a response to the studies I've cited would require its being widespread.

6. Rosenthal on Metacognition

Rosenthal has commented on various metacognitive measures, including confidence, and some of those comments might be used in defence of HOT. It will be helpful in this section to adopt Rosenthal's terminology and distinguish states of higher-order awareness (HOAs), which he takes to be sufficient for consciousness, from the broader category of HOR or metacognitive state. At points he has seemed to implicitly agree that c-states are HOAs, for he cites with approval studies that claim to locate the neural basis of HOAs in the prefrontal cortex, and these studies use confidence ratings as their behavioural measure of HOAs (Rosenthal, 2011, p. 367, Figure 1).[12] Elsewhere he seems to explicitly refer to confidence as a HOA (2012, p. 1429). But other comments are clearly opposed to the view that c-states are HOAs, and these I should address. He has argued that metacognition in general is distinct from higher-order awareness and has made similar comments about confidence in particular. Here I will consider these comments and the arguments, sometimes implicit, that might be constructed from them.

6.1. Differences between metacognition and HOAs

Rosenthal (2012) suggests that there are some general differences between HOAs and metacognitive states, which is presumably meant to suggest that they also differ in their consciousness-making properties. Rosenthal holds that most HOTs are unconscious, since for them to be conscious requires a third-order HOT, and he would like to avoid an infinite regress. Metacognitive states, he claims, are

12 One could imagine a new line of argument for HO that does not require TP, relying instead on brute correlation between consciousness and activity in brain areas thought responsible for HOAs. However, the studies cited by Lau and Rosenthal (2011) in support of this claim have been called into serious doubt by Jannati and Di Lollo (2011) and Bor *et al*. (2017).

generally conscious, which suggests that these are different types of states. The premise regarding metacognition is false. In the empirical studies I have presented, confidence is presumably conscious because it is measured by subjects' reports. But researchers often study unconscious confidence using wagering paradigms (e.g. Dienes and Seth, 2010). Indeed, the upshot of the usefulness of the wagering paradigm is that much of the time metacognition is *unconscious*. And Rosenthal's inference is invalid as well. Some chairs may be mostly wood and partly metal and others mostly metal and partly wood, but both are chairs, nonetheless. Likewise with HOAs.

Rosenthal (2012) offers a similar, though not quite parallel, argument from utility. Metacognitive states, Rosenthal claims, have utility in so far as they allow us to fine-tune our first-order judgments. Conscious states, on the other hand, have no utility. The argument here is valid, since any utility would be sufficient to disqualify a mental state as conscious for Rosenthal. Of course, Rosenthal's claim that conscious states lack utility is quite controversial and could be challenged. But confidence needn't be conscious, so there is no obvious conflict. Confidence *is* conscious in the studies I've presented, so I suppose Rosenthal's claim would entail that in those particular studies confidence ratings couldn't possibly count as metacognition. But that claim flies pretty outrageously in the face of scientific consensus. We have more reason to accept that confidence is a metacognitive measure than we do the claim that consciousness is useless.

6.2. More on TOTs

Having denied that tip-of-the-tongue experiences (TOTs) are HOAs on the basis that their intentional objects are in the future, Rosenthal (2012) considers another interpretation — TOTs target beliefs that exist in the present but are currently inaccessible to consciousness. Rosenthal argues that these HORs do not entail consciousness because they target that content in a specific way. They describe the object mental state as containing the answer to a specific question, but don't specify what the answer is. He concedes that HOTs do not capture all the content of the states they target either but argues that the TOT is different in that it 'leaves out too much' (*ibid.*, p. 1428). What counts as too much? He says that 'one would regard the state as conscious only if one were aware of the state in respect of that aspect of the

content that is of current interest' (*ibid.*, p. 1428). So being of current interest seems to be the criterion.

For my purposes this argument about TOTs is neither here nor there, but Rosenthal's response resembles one I will consider later regarding confidence, so his comments here will be relevant later. For now, I will simply note that being of current interest is likely not a sustainable criterion for which properties a HOR makes conscious. The thought 'I am perceiving a red square' will not capture every property of the perceptual event it makes conscious. It cannot for the familiar reason that thought is digital and vision is analogue. I perceive the size, contrast, and many other properties of the square, even if these are not of current interest or mentioned in the HOT. When I return to this issue, however, I will consider a possible version of HOT which would try to deny this.

6.3. Rosenthal on confidence

When Rosenthal does consider confidence specifically, he treads a fine line between conceding that confidence ratings are HORs with some important relation to HOAs, while denying that confidence constitutes a HOA. Much of what he says on the subject is not aimed directly at my line of argument, which is to show that c-states *just are* HOAs. Instead he is largely concerned with the claim that confidence is a reliable indicator of the presence of a HOA. This interpretation does exist in the literature. Some studies of metacognition use confidence as a measure of conscious awareness, the assumption being that one is confident in what one perceives because one is aware of it. The studies I have presented as evidence against HO do not assume that confidence is contingent on consciousness, instead taking separate measures of type-1 performance, type-2 performance, and consciousness. Since the three are shown to dissociate, this is the better interpretation. Rosenthal (2019) explicitly notes that consciousness and confidence dissociate, and on this point we agree. But there is one passage that could be interpreted as targeting my line of argument.

Rosenthal considers the claim that being aware of something typically (though not always) 'results in' or 'carries with it' confidence about it (*ibid.*). He seems to avoid specifying whether the relation is causal or constitutive, but if confidence constitutes a HOA and confidence and consciousness dissociate, HOT is sunk. So let's assume the causal relation. Rosenthal posits a series of such relations. The HOA involved in a conscious perception 'carries with it'

confidence that one has the perception, which in turn 'brings with it' confidence that the perception is accurate, which then 'is' confidence 'about' the stimulus represented in perception. The final step appears to be an identity relation, but it also takes the chain one step further than the studies I have presented. As I have noted, researchers take much care to ensure that confidence ratings are not targeting the stimulus itself, but the perceptual state. But Rosenthal might make something of the claim that HOAs cause two sorts of c-states — confidence that one is perceiving, say, a square, which in turn causes, or possibly constitutes, confidence that one's perception is accurate. Rosenthal then addresses the possibility that confidence ratings can be taken as reliable indicators that a HOA obtains, in which case their dissociation from consciousness could refute HO.

First, he points out that if the perception itself is degraded it might not produce confidence in the perception. One will be confident that one is having a perception, just not confident that one is perceiving a square. Or one might choose 'square' in the type-1 task but not be confident that this is accurate. In the studies I cite in support of my claim, the stimuli *are* sometimes degraded. But the trials that I claim are counter-examples to HO are not trials in which one therefore lacks confidence. They are trials in which subjects are correct on the type-1 task, have high confidence, and yet the perceptions fail to be conscious. So Rosenthal's point here is irrelevant to my argument.

Rosenthal's next objection, however, comes closer to touching my argument. The argument involves type-2 blindsight. Type-2 blindsight has been described in various ways. Sometimes it is described as a sort of 'shadowy' perceptual experience of an object, accompanied by confident and accurate judgments that there is an object there (Brogaard, 2015). Rosenthal describes it as a total lack of visual experience of the object, accompanied by confidence that there is an object. Thus, there can be confidence about a stimulus without 'confidence in the perception'. He then claims that something of the sort probably occurs in normals, and this shows that one can have confidence without HOAs. That is, since Rosenthal has posited a causal route from the HOA down through two types of confidence, if a form of confidence further down the causal path can occur in the absence of a form of confidence earlier in that path, this shows there is an independent route to confidence that does not involve the consciousness-making HOAs, and thus that confidence is not a reliable indicator of the presence of HOAs. It would then be no threat to HOT if confidence were shown to dissociate from consciousness.

Of course, I do not subscribe to this causal analysis of the relation between confidence and HOAs. I claim that confidence *is* a HOA because it satisfies TP. Confidence is an intentional state whose object is a mental state to which it is related in the right way. But even if you do buy the causal analysis, Rosenthal's objection relies on a particular interpretation of the blindsighter's experience. On some interpretations, the type-2 blindsighter does have a perception, just a very hazy one. In this case the confidence that there is a stimulus is based on a perception and the posited causal route remains intact. But more importantly, Rosenthal's argument begs the question. His reason for claiming that no HOA obtains in type-2 blindsight is that the subject is not conscious of the stimulus. But this assumes HO.

Finally, there is Rosenthal's (2011) response to Kanai, Walsh and Tseng (2010). Kanai *et al.* show that in cases of attentional blindness, where a stimulus is present but subjects report no conscious experience of it, subjects will report low confidence in this judgment, suggesting that there is 'sensory metacognition' in the absence of consciousness. Rosenthal claims that, even though such a metacognitive state may obtain, it differs from the HOA, which in such cases may have the content that there is no perception.

> Even though sensory metacognition and awareness are related, a specific kind of higher-order representation is required on the higher-order view for conscious awareness to occur… It is entirely compatible with the higher-order view that subjects do not consciously perceive when they have a higher-order representation of not perceiving, and this accounts for the metacognitive ability. (Rosenthal, 2011, p. 64)

But, as we have seen, Rosenthal has yet to establish that confidence doesn't constitute awareness of the right sort.

An anonymous referee suggests that Rosenthal's general line of argument in this passage, and against the identification of any metacognitive state with a consciousness-making HOR in general, is the following:

> Metacognition invariably involves assessments of the accuracy of first-order mental processes (e.g. how confident are you in what you saw — that is, how accurate was your perception?). But the kind of HO awareness that explains consciousness, on Rosenthal's approach, does not involve such assessments. To be in a conscious state, on the HO view, is just to be aware of yourself as being in a state in a specific kind of way. Your awareness does not make 'commentary' about your FO state — it just says something like, 'I am in state X'. And it is this, as well as several other differences, that explains why consciousness and metacognition (as it is standardly studied) come apart. In particular, the fact

> that one metacognitively judges a perception to be accurate to whatever degree does not assert that one, oneself, is currently in that perceptual state — and it is only the latter kind of awareness of a state that engenders consciousness.

I am not entirely sure what to make of the claim that the fact that one 'metacognitively judges a [presumably one's own] perception to be accurate… does not assert that one, oneself, is currently in that perceptual state' unless it amounts to the old temporal objection. If I judge my *current* perception to be accurate, surely this just is a judgment that I am in a perceptual state. What differs is the property attributed to the perception. In one case I judge that I am in a, say, square-perceiving state and in the other I judge that I am in an accurate perceptual state. If this is the point, then it is reminiscent of Rosenthal's comments on TOTs, where he claims that whether a HOR will entail consciousness depends on whether it picks out the property of current interest. It's not clear that this is Rosenthal's considered view, but we can explore something along these lines. Suppose a HO theorist claims that a HOR makes conscious precisely those properties of the first-order state that occur in the HOR and no more. If the HOR only mentions the square, then we are not conscious of its size or colour. Likewise, if the HOR only mentions accuracy then only accuracy will be conscious. Of course, often the accuracy isn't conscious even when confidence is high, as revealed by wagering paradigms, so this response is still problematic for HOs sufficiency claim. And there is the further problem that these attributions of accuracy are themselves typically accurate. This means that there is a HOR of the percept's nature, not just of its accuracy. That is, in order to reliably decide whether my percepts accurately track the properties of a stimulus, I must have information about the properties of the stimulus and information about the properties of the perception. If they match, I can judge the perception as accurate. This means that an accurate representation of accuracy implies a representation of the property in question, and thus, for HO, a conscious percept.[13]

[13] A potential line of response could borrow from Proust (2015), who makes a distinction between the proximal and distal intentional object of a metacognitive state. If the proximal object is some heuristic, say, reaction time, then confidence can accurately represent the distal object, the perception's accuracy, in virtue of the correlation between reaction time and accuracy. But this account is unsupported by our best models of confidence (Pleskac and Busemeyer, 2010), which rule out such heuristics, instead positing direct access to the first-order signal as the basis of confidence.

In this section I have addressed an assortment of comments by Rosenthal on the relation between metacognition in general, confidence in particular, and higher-order awareness. I have argued that his various attempts to distinguish the two are not sufficient to save HO in the face of empirical falsification.

7. Conclusion

I have argued that empirical evidence shows that higher-order theories of consciousness are false. HO predicts that consciousness and metacognition should correlate, but this is not the case. I have then argued that HO theorists will be unable to reinterpret these studies in any way that could save HO. While I lack the space to explore every facet of every version of HO, the points I have made are general enough that they should apply quite broadly. I applaud HO theorists for making bold empirical predictions that make their philosophical theory of consciousness falsifiable. This is progress. As it turns out, the evidence does falsify HO, but this is also progress. Philosophy is getting somewhere.

Acknowledgments

Thanks to Matthew Fulkerson, Jonathan Cohen, Rick Grush, William Bechtel, and the anonymous referees for their comments on earlier drafts of this article. Thanks also to Letizia Ragusa, Donna Balderrama, David Pitt, Michael Shim, Mark Balaguer, Foad Dizadji-Bahmani, and Talia Bettcher.

References

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

Block, N. (2011) The higher-order approach to consciousness is defunct, *Analysis*, **71** (3), pp. 419–431.

Berger, J. (2014) Consciousness is not a property of states: A reply to Wilberg, *Philosophical Psychology*, **27** (6), pp. 829–842.

Bor, D., Schwartzman, D.J., Barrett, A.B. & Seth, A.K. (2017) Theta-burst transcranial magnetic stimulation to the prefrontal or parietal cortex does not impair metacognitive visual awareness, *PLoS One*, **12** (2), e0171793.

Brogaard, B. (2015) Type 2 blindsight and the nature of visual experience, *Consciousness and Cognition*, **32**, pp. 92–103.

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

Brown, R., Lau, H. & LeDoux, J.E. (unpublished) The misunderstood higher-order approach to consciousness, *ArXiv*, [Online], https://psyarxiv.com/xpy8h/.

Carruthers, P. (1996) *Language, Thought and Consciousness*, Cambridge: Cambridge University Press.

Carruthers, P. (2004) HOP over FOR, HOT theory, in Gennaro, R. (ed.) *Higher-Order Theories of Consciousness*, pp. 115–135, Philadelphia, PA: John Benjamins.

Carruthers, P. (2011) *The Opacity of Mind*, New York: Oxford University Press.

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

Chisholm, R.M. (1950) The theory of appearing, in Black, M. (ed.) *Philosophical Analysis*, pp. 102–118, Ithaca, NY: Cornell University Press.

Cleeremans, A. (2011) The radical plasticity thesis: How the brain learns to be conscious, *Frontiers in Psychology*, **2**, art. 86.

Coleman, S. (2015) Quotational higher-order thought theory, *Philosophical Studies*, **172** (10), pp. 2705–2733.

Dienes, Z. & Seth, A. (2010) Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task, *Consciousness and Cognition*, **19**, pp. 674–691.

Galvin, S.J., Podd, J.V., Drga, V. & Whitmore, J. (2003) Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions, *Psychonomic Bulletin & Review*, **10** (4), pp. 843–876.

Gennaro, R. (2012) *The Consciousness Paradox*, Cambridge, MA: MIT Press.

Jachs, B., Blanco, M., Grantham-Hill, S. & Soto, D. (2015) On the independence of visual awareness and metacognition: A signal detection theoretic analysis, *Journal of Experimental Psychology: Human Perception and Performance*, **41** (2), pp. 269–276.

Jannati, A. & Di Lollo, V. (2011) Relative blindsight arises from a criterion confound in metacontrast masking: Implications for theories of consciousness, *Consciousness and Cognition*, **21**, pp. 307–314.

Kanai, R., Walsh, V. & Tseng, C. (2010) Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness, *Consciousness and Cognition*, **19**, pp. 1045–1057.

Kant, I. (1781/1996) *Critique of Pure Reason*, Indianapolis, IN: Hackett.

Koriat, A. (2000) The feeling of knowing: Some metatheoretical implications for consciousness and control, *Consciousness and Cognition*, **9**, pp. 149–171.

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

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

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

Lycan, W.G. (2001) A simple argument for a higher-order representation theory of consciousness, *Analysis*, **61** (1), pp. 3–4.

Lycan, W.G. (2004) The superiority of HOP to HOT, in Gennaro, R. (ed.) *Higher-Order Theories of Consciousness*, pp. 93–113, Philadelphia, PA: John Benjamins.

Maniscalco, B. & Lau, H. (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings, *Consciousness and Cognition*, **21**, pp. 422–430.

Pleskac, T. & Busemeyer, J. (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence, *Psychological Review*, **117** (3), pp. 864–901.

Proust, J. (2015) *The Philosophy of Metacognition*, New York: Oxford University Press.
Rosenthal, D. (1986) Two concepts of consciousness, *Philosophical Studies*, **49**, pp. 329–359.
Rosenthal, D. (1997) A theory of consciousness, in Block, N., Flanagan, O. & Güzeldere, G. (eds.) *The Nature of Consciousness: Philosophical Debates*, pp. 729–753, Cambridge, MA: MIT Press/Bradford Books.
Rosenthal, D. (2005) *Consciousness and Mind*, New York: Oxford University Press.
Rosenthal, D. (2011) Exaggerated reports: Reply to Block, *Analysis*, **71** (3), pp. 431–437.
Rosenthal, D. (2012) Higher-order awareness, misrepresentation, and function, *Philosophical Transactions of the Royal Society, B*, **367**, pp. 1424–1438.
Rosenthal, D. (2019) Consciousness and confidence, *Neuropsychologia*, **128**, pp. 255–265.
Rosenthal, D. & Weisberg, J. (2008) Higher-order theories of consciousness, *Scholarpedia*, **3** (5), art. 4407.
Russell, B. (1912/1997) *The Problems of Philosophy*, New York: Oxford University Press.
Sauret, W. & Lycan, W.G. (2014) Attention and internal monitoring: A farewell to HOP, *Analysis*, **74** (3), pp. 363–370.
Schwitzgebel, E. (2015) Belief, in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*, Summer 2015 edition, [Online], https://plato.stanford.edu/archives/sum2015/entries/belief/.
Wilberg, J. (2010) Consciousness and false HOTs, *Philosophical Psychology*, **23** (5), pp. 617–638.

Paper received December 2019; revised April 2020.