The merits of higher-order thought theories

Sam Coleman[1]

Abstract: Over many years and in many publications David Rosenthal has developed, defended and applied his justly well-known *higher-order thought theory* of consciousness.[2] In this paper I explain the theory, then provide a brief history of a major objection to it. I suggest that this objection is ultimately ineffectual, but that behind it lies a reason to look beyond Rosenthal's theory to another sort of HOT theory. I then offer my own HOT theory as a suitable alternative, before concluding in a final section.

Keywords: Consciousness. Higher-order thought. Mental quotation. Misrepresentation. Acquaintance

1 Rosenthal's higher-order tought (HOT) theory

The main question HOT theory aims to answer is what makes a mental state conscious. There is good evidence, deriving both from everyday experience and empirical science, that there exist unconscious mental states. Even setting aside that thesis for a moment, there are clearly many states of one's brain and body which are not conscious—e.g. those corresponding to the growth of neuronal connections, or to the activity of enzymes in the gut. Therefore it is of interest to understand what makes the difference between such states and conscious states. What it means to say that a state is not conscious is that it is in no way *felt by the subject*. Rosenthal parses consciousness in this sense as *awareness*: to say a state of the subject is conscious is to say that the subject is aware of it. State consciousness is thus equivalent to the subject's consciousness of the state. This equation between a state's consciousness and a transitive relation of subjective awareness of the state opens the door to the HOT account.

[1] Reader in Philosophy, University of Hertfordshire, Hatfield – United Kingdom. E-mail: s.coleman@herts.ac.uk

[2] See Rosenthal (2005) for a collection of some of the most important of these, in updated form, along with some excellent new material.

http://dx.doi.org/10.1590/0101-3173.2018.v41esp.04.p31

On that account, what a mental state's consciousness consists in is the subject's having an accompanying thought to the effect that she is in that state. Say that Paulo undergoes an injury while playing soccer. It is possible that the pain this injury causes does not become conscious while Paulo is playing, for his system may suppress it while he strives for his team. But with the match over Paulo begins to feel the pain—it becomes conscious. According to HOT theory, what makes the difference to the conscious instantiation of the pain is that Paulo now comes to have a thought that represents that he is in just such a pain state. Since the pain is a mental state, in the broad sense in which sensory contents count as mental, and the thought, another mental state, represents it, the thought can be termed a *higher-order* mental state, specifically a *higher-order thought*. On HOT theory the pain's consciousness—its being felt—involves nothing other than the com-presence of such a HOT with the pain.

Not just any kind of thought representing the pain suffices for pain *sensation*, however. Obviously the thought must overlap in time with the pain—thinking about a pain he experienced playing soccer aged five need not cause Paulo to feel it again. The thought's content must also be *assertoric*—it must affirm that Paulo is in the relevant pain state. A hesitant or doubting HOT will hardly generate the subjective impression that one is in pain. Moreover, importantly, the awareness the thought provides of the pain must not seem to the subject to be mediated in any way. One can acquire a belief that one is in pain (or other mental state) by inference, or through being informed by a psychological or physiological expert. But in such cases one need not come to feel the state in question, on that basis alone. By contrast, conscious awareness of a mental state has a striking *apparent directness*. The thoughts that provide state consciousness, for Rosenthal, are not the result of inference. Nor, in having them, is one aware of any intermediary between oneself and the conscious state in question. The apparent directness of consciousness is neatly captured on Rosenthal's model by the fact that HOTs are typically unconscious: since one is unaware of that by means of which the target state is made conscious, one's awareness of it will not seem mediated. So the overall picture on HOT theory is that the right kind of unconscious HOT, when directed at a mental state, makes the subject conscious of that state.

Rosenthal's theory is a highly elegant, substantive, and fruitful account of that in which a mental state's consciousness consists. It combines simplicity with considerable power, and, operating at a cognitive level of analysis, paves

the way for a functionalist reduction of consciousness to physical processes in the brain.

2 An objection to HOT theory

This sophisticated development of a venerable theory of consciousness[3] has faced several objections, perhaps the most important of which is the *mistargeted representation problem*,[4] influentially aired by Neander (1998).[5] I will focus on it, and will suggest that while in existing iterations it does not provoke serious problems for Rosenthal, it nonetheless does bring to our attention a possible reason for seeking an alternative to his HOT theory.

Consider a visual experience as of a red London bus. In the good case, a red London bus-ish percept is present even while a HOT suitably represents one as being in such a perceptual state. But percept and HOT are wholly separate states for Rosenthal, which creates the possibility of two sorts of slippage between them. In *misrepresentation* the percept is unfaithfully represented: e.g., the HOT represents the subject as in a *blue* London bus-ish perceptual state. In *targetlessness* the percept is non-existent, but still the HOT represents the subject to be in such a red London bus-ish perceptual state. Rosenthal's challenge is what to say about such cases. Take misrepresentation: it seems he could either say the subject experiences as of a red London bus, or as of a blue London bus. He cannot apparently deny that *any* conscious state results from this combination of HOT and percept, since by hypothesis such combinations suffice for conscious states. If he gives the former response, the objection goes, one wonders what the point is of HOTs in *HOT theory*. But if he offers the latter response, it seems to make the perceptual state redundant as regards consciousness. As for targetlessness, the question is whether a conscious state obtains in the absence of the perceptual state, with similarly dilemmatic options awaiting Rosenthal in reply. However he responds, thus, Rosenthal seems to be left in the position of maintaining that one component of his two-component HOT theory is surprisingly under-involved in the production of consciousness.

[3] Locke and Aristotle, to name but two, held something similar. See Caston (2002).

[4] See also the 'rock' problem (GOLDMAN, 1993; STUBENBERG, 1998).

[5] See also Byrne (1997).

In reply, Rosenthal has been clear about his preferred treatment of such cases. He explicitly entrusts HOTs with generating the 'subjective mental appearances' that comprise conscious episodes. His thinking is somewhat as follows. A conscious mental state is, after all, one the subject is aware of herself as being in. Since the best way to construe this awareness is in terms of representation, Rosenthal argues, this means some mental state of the subject represents her to be in the relevant mental state she is aware of herself as instantiating. For instance, a HOT represents her to be in a red London bus-ish visual state. But, once HOTs are installed as the providers of subjective mental appearances, why should it then matter whether the target state exists? One can surely be aware of oneself *as* being in a mental state even if there is no such state, just as one can be aware as of a unicorn before one without any need of a unicorn. How then does it *feel* to be aware of oneself as instantiating a red London bus-ish percept when no such percept obtains? Conscious mental life is the stuff of appearances, Rosenthal emphasises: thus, since the appearance to the subject is the same whether or not she tokens a red London bus-ish percept (assuming the same HOT is in place), her experience is indistinguishable across the two cases. In a similar way, perceptual hallucinations can be indistinguishable from their veridical counterparts. Corresponding remarks apply to misrepresentation: HOTs govern the subjective mental appearances constitutive of the subject's stream of consciousness.

Many commentators have felt Rosenthal's treatment of mistargeting to be unsatisfactory. Their shared misgiving is that lower-order states, those successfully represented in good cases, do not ultimately seem to matter very much to the subject's stream of consciousness. Block, in a recent critique, has attempted to focus these rays of dissatisfaction into a precise objection. First airing the suspicion that Rosenthal's approach ends up 'jettisoning the first-order content as constitutive of consciousness',[6] Block proceeds to allege an outright incoherence in HOT theory, due to Rosenthal's "holding both that an appropriate higher-order thought is sufficient for a conscious state and that being the object of an appropriate higher-order thought is necessary for a conscious state". (BLOCK, 2011, p. 443).

A targetless HOT supports a conscious state, by the sufficient condition. Yet since it lacks any target, and we can stipulate that it is not targeted by another higher-order thought, it follows that a conscious state

[6] I will argue (§3) that Block's suspicion is justified, though I spell it out in a different way, and reject his specific objection below.

obtains without anything being the object of an appropriate HOT—violating the necessary condition. Block concludes that HOT theory's necessary and sufficient conditions conflict, making it incoherent.

But Rosenthal's theory is not really vulnerable to this objection. There are at least two possible escape routes open to him.[7] First, he might reject the sufficient condition: Rosenthal need not accept that a lone HOT suffices for a conscious state. In order for a conscious state to obtain, he might say, a lower-order state must be appropriately represented by a HOT. For it is the lower-order state that becomes the conscious state, if any. But in Block's scenario no lower-order state exists, hence no conscious state can obtain. What would obtain instead, admittedly, is a *mental appearance as of an existent conscious state*, which Rosenthal might rule to be indistinguishable, from the subject's perspective, from there actually existing such a state. But, as previously noted, mental life for Rosenthal is the stuff of appearances, and just as susceptible to misrepresentation, or hallucination-type states, as is perception. Thus Rosenthal might find nothing problematic in this possibility.

His actual reply, (ROSENTHAL, 2011) and the second escape route, is more subtle. Rosenthal accepts Block's construal of his theory in terms of the stated necessary and sufficient conditions. But he argues that the conscious state—the state of which the HOT makes one aware—need not be an *existent* state. In such cases it will be true that being represented by a HOT suffices for a conscious state. And such higher-order representation is also necessary for a conscious state to obtain. But, since, to repeat, conscious mental life is the stuff of appearances, if a HOT successfully supplies the appearance to the stream of consciousness as of an in fact non-existent first-order state, then a conscious state indeed obtains—only it is not one that the subject is truly *in*, even though she is aware of herself *as* being in it. One can see that this is simply another iteration of the reply Rosenthal made to the Neander-style objection above.

Rosenthal's response does however embroil him in a further dialectic, concerning whether one can undergo the 'subjective appearance' as of being in a conscious state without actually being in such a state—i.e. with whether such subjective appearances in fact suffice for an existent state of consciousness. Block himself complains that Rosenthal hereby implausibly reduces conscious states to 'intentional objects'—mere representata, which could not possibly 'matter' as

[7] See Coleman (2015) for a third, which involves partial rejection of the necessary condition.

much as existing states (BLOCK, 2011b, p. 444).[8] Kriegel (2009, p. 131) attempts to make this complaint more concrete, by offering it as an 'obvious truism' that "[t]here's something it's like for the subject only if she is in a conscious state."

Underlying this Block-Kriegel line of objection seems to be the presumption that in first-person life, no appearance/reality gap is possible.[9] If one starts from this presumption then the appearance of being in a conscious state must co-incide with actually being in one. And that will put pressure on Rosenthal in cases of targetlessness, where he wants to say that there is every subjective impression of one's being in a conscious state that one is not in fact in. But, to note once more, Rosenthal makes a point of rejecting the appearance/reality coincidence for first-person life. He is strongly concerned to maintain that mental life is just another area of *life*—and in life appearance and reality can and do come apart. That leaves his response to Block intact, and leaves him untroubled by Kriegel. There is of course a further debate to be had about the status of the mental appearance/reality principle. But that is a substantive issue, and the outcome of the debate hard to foresee. Therefore a presumption of its truth cannot be the starting point of any objection to Rosenthal's HOT theory. In this light it can be seen that the Block-Kriegel line of objection actually begs the question.[10]

In this brief survey of a major line of objection to HOT theory I have suggested, effectively, that nobody has been able to lay a glove on Rosenthal. Nonetheless, I myself share the broad misgivings of the commentators. While I do not believe any knockdown objection to HOT theory can be derived from the phenomena of mistargeting, I will explain below that Rosenthal's treatment of these phenomena does open our eyes to what might appear on reflection an undesirable feature of his theory. If that is correct, it will give us reason to look for an alternative kind of HOT theory that lacks the undesirable feature. I will proceed to offer such a theory in §4.

3 Does Rosenthal's HOT theory fulfill its promise?

Here is a natural way to capture the naive appeal that Rosenthal's HOT theory has had for many people as an analysis of consciousness. Consider a

[8] . Yet many non-existent things matter a great deal—e.g. Brexit.

[9] Oft-maintained—see Kripke's (1972) influential use of this principle to argue against physicalism.

[10] For his part Rosenthal (2011) adverts to phenomena such as dental fear as evidence that the appearance/reality principle for consciousness is false.

red London bus-ish percept, but one of which the subject is, for whatever reason, currently not conscious. Perhaps it is a state that does not enjoy focal attention, because the subject is engaged in an absorbing task, and that suffices for it to lack consciousness.[11] Or perhaps the visual state is lodged in a *blindsighter's* brain. Though she may glean much information from it, even being able to report on the stimulus's colour when prompted, something is clearly missing for this subject—that would be even were she one of Block's mythical *superblindsighters* (BLOCK, 1995). What's missing, of course, is that she is not *conscious* of the red bus-ish percept: in other words there is nothing it is like for her to token it.

HOT theory seems to offer a pleasingly substantive account of the extra something needed to get this visual percept into the subject's stream of consciousness, making her conscious of its qualities—e.g. colour qualities. The answer is that the red bus-ish percept would enter her conscious mental life just in case it became suitably represented by a HOT.

But we have now been alerted, by Rosenthal's treatment of mistargeting, that HOTs call the shots regarding the contents of consciousness. What gets into the subject's stream of consciousness is all and only what her HOTs represent as being the case. So let us imagine that our blindsighter has a (currently impossible) operation to undo her blindsight, reinstating visual consciousness in the relevant area of her visual field. On Rosenthal's scheme, this operation, let us imagine, makes her capable of tokening the HOTs requisite for visual consciousness of items presented to the hitherto blind field.[12] Post-operation, things proceed promisingly: her red London bus-ish percept is suitably HOT-targeted, and she becomes (happily!) conscious as of a red London bus. But then something malfunctions, and, while its corresponding HOT continues to operate, the red London bus-ish percept itself is somehow destroyed. Does the subject notice any change? It seems clear that on Rosenthal's account she should still be experiencing red London bus-ishly: for it is HOTs and the presentations they make that determine conscious mental life, and this HOT

[11] As with Armstrong's driver (1981).

[12] It is unclear quite how blindsight functions, so unclear the extent to which this HOT-style diagnosis of a blindsight visual state's lack of consciousness in fact applies. Nonetheless some accounts implicate intra-visual cortex re-entrant processing in the awareness mechanism for conscious states, and if this is affected in blindsight, such a diagnosis becomes extremely tempting (but see Lau 2008 for an alternative HOT-based account of blindsight). If unhappy with my use of blindsight, which I select for vividness, the reader should focus on the more conventional case of a merely unconscious visual state, i.e. one missing a corresponding HOT.

remains in service, dutifully representing the presence of a red London bus-ish visual percept. So far, this is only another illustration of Rosenthal's treatment of mistargeting: In our scenario we have simply gone from 'veridical' conscious experience (accurate representation of a mental state) to a targetless case, and what Rosenthal says about this latter species applies.

Yet it is evident that, once it has disappeared, the visual percept makes no contribution to the stream of consciousness. The subject cannot be conscious of the red bus-ish percept any more, since it no longer exists.[13] If one is genuinely conscious *of* x, as opposed to being conscious merely *as of* x, then x certainly exists. And Rosenthal's theory—on the natural way of explicating its appeal above—was supposed to account for our genuine consciousness of mental states, not merely our consciousness as of mental states. We wanted to know what it would take to make our blindsighter conscious of her red London bus-ish percept—what was needed for *that very* visual percept to be like something for her; not merely what it would take to give her *every impression* that *some such* percept was like something for her.

As noted, in the present case the disappearance of the visual percept makes no difference to the subject's conscious mental life—how things seem subjectively to her—since the HOT keeps supplying the relevant appearances to her stream of consciousness. The complaint, then, is not that Rosenthal's theory permits misrepresentation, nor that it allows subjects to experience in the absence of first-order states, nor even that first-order states are redundant. The point, rather, as this case illustrates, is that the subject simply *never becomes conscious of the first-order state at all, even in the good case.*

An analogy should help to clarify and substantiate this claim. Imagine the following situation, whose odd structure will soon acquire obvious significance. A square canvas, painted wholly purple, is hanging in a gallery. A micro-camera is facing the picture at head-height, suspended in mid-air, and so tiny as to be practically invisible to the naked eye. The camera is sending its feed to a projector, affixed to the wall opposite the picture, and the projector is projecting a faithful image of the purple canvas *onto* the surface of the canvas itself. This setup already sounds somewhat high-tech and futuristic, so let us add that the projector projects not a flat 2-d image but a fully life-like 3-d holographic image. Now someone approaches the picture and takes up a position next to (but not touching) the camera, and, conveniently, not

[13] Cf. the time-overlap condition in the account of HOT theory on p. 32.

blocking the stream of light from the projector. The observer notices neither camera nor projector—again conveniently, for our purposes. As a result, the subject observing the 3-d projection takes herself to be seeing the purple canvas—that is how things seem. Indeed she is well, and reliably,[14] informed about that canvas by her visual experience. She goes away none the wiser and returns next day. Overnight, however, the canvas is stolen in a philosophical prank. But, fortunately, the feed from the painting has been stored by the micro-camera, and can continue to be streamed, on loop, to the projector. The gallery will be in trouble if the canvas's theft is discovered, and hope to delay that circumstance while they ascertain its whereabouts. The holographic projection is so good, so visually 'solid', that the visitor, and other visitors, notice nothing, and continue to take themselves to be seeing the canvas.

On the second day the painting is *not there*, hence the visitor cannot be seeing it. But, she is seeing exactly what she saw on the previous day, namely: a holographic projection from the projector. Therefore, *she did not see the painting on the first day*, *either, even when it was present*. Despite being an accurate and reliable representation of its target, the projection in fact *screened the painting off* from the visitor.

Corresponding remarks apply to the structure of a conscious state on Rosenthal's HOT theory. The fabric of a Rosenthalian HOT is simply too 'thick' for first-order targets ever to penetrate to the subject's stream of consciousness. Managing entirely the content of one's stream of consciousness, HOTs effectively block anything else getting in. At best we consciously experience veridical echoes of first-order states; but of the first-order states themselves we are not conscious. This manifests in the possibility that, on their removal, we are left with just the conscious impression they made when extant, and experience no change. *Experiencing no change* means that the same appearances are contributed to the stream of consciousness as were contributed while the relevant first-order state persisted and was HOT-targeted: it is important to note that it is not that we are *oblivious* to some change in the appearances, it is that there is *in fact* no change to the appearances (regardless how the subject judges). But, since the relevant appearances supplied to the stream of consciousness remain intact with the first-order state removed, it follows that the latter was contributing no appearances to subjective mental life even while it was present. It is not as if the HOT must now '*pick up some slack* '

[14] Since the projector's representation of the canvas is causally connected in useful ways, via the camera, with the presence and properties of the canvas itself.

in generating appearances formerly due to the first-order state. Rather, for Rosenthal, it has been the HOT governing appearances added to the stream of consciousness all along. In other words, we were not ever quite conscious of the first-order state. Rosenthal comes close to acknowledgement of this verdict, averring that a first-order state "can contribute nothing to phenomenology apart from the way we're conscious of it." (ROSENTHAL, 2004, p. 32). That 'way', of course, is precisely the HOT that represents the state. Hence, all that a first-order state contributes to the stream of consciousness is a proxy that represents it; otherwise put, it directly contributes nothing. *It* does not appear to consciousness. Therefore, on HOT theory, we are simply not conscious of our first-order states.[15]

Rosenthal's theory is far from being incoherent. But on examination it turns out to lack the appeal it initially appeared to possess. It does not account for the consciousness of first-order states, since we are never strictly conscious of them. There is nothing *they* are like for us, given Rosenthal's envisaged structure for conscious states. This constitutes not a refutation, but a clarification, of his theory. Rosenthal's primary concern is with what supports the stream of consciousness, and nothing I have said puts his theory in doubt regarding that objective [see Brown (2012)]. Still, we are within rights to seek a theory that does hold the appeal we wrongly glimpsed in Rosenthal's theory. To adopt a term of Kriegel's, we might hope for a theory that allows us genuine *intimacy* with our first-order states, in consciousness.

4 QHOT THEORY

When he makes his (unsuccessful) allegation of an incoherence in Rosenthal's HOT theory, Block diagnoses the underlying problem as being that 'in the HOT account…content is doubled'—i.e. qualitative content is associated both with the HOT, which represents the first-order state, and with the first-order state itself, which (typically) represents some worldly situation (BLOCK, 2011, p. 447).[16] Rosenthal's decision to place the provision of subjective mental appearances exclusively in the hands of HOTs results, I argued, in a screening-off effect: the qualitative content of first-order states never makes it to consciousness. On this Blockian diagnosis,

[15] I do not imply that *HOTs* are conscious: though they supply appearances to the stream of consciousness, these are appearances not of themselves, but as of first-order states.

[16] cf. Kriegel's (2009) 'qualitative' vs. 'schmalitative' contents.

what would seem required for the requisite intimacy with our first-order states—to make them conscious—is the removal of one layer of qualitative content, namely that pertaining to the HOT. Closing his critique, Block recommends a 'pointer theory', as advocated by Lau, on which 'a higher-order probabilistic "pointer" *that has no content of its own* refers to a first-order representation", such that "The consciousness of seeing a red square could be glossed in language as: Probably (red square at 3:00)." (BLOCK (2011, p. 447, my emphasis; LAU, 2008).

Block is in my view right about this promising direction of travel for HOT theories. However, he is wrong to place his hopes in Lau's theory in particular. For on Lau's theory, as his ingenious treatments of blindsight and 'hallucination' make clear, no less than on Rosenthal's theory, representations provided by HOTs actually *constitute* the subjective stream of consciousness. Thus Lau's theory will also produce the unwelcome screening-off effect described above.[17]

I will now present my own HOT theory, which aims to provide an account that prevents screening off, making room for genuine consciousness of our first-order states.

Kriegel explains the notion of a 'display sentence', via the example of a bridge with 'under construction' painted on it.[18] Crossing, you think to yourself 'This bridge is under construction'. Kriegel suggests that you actually have before you the complete sentence 'This bridge is under construction', with the subject term supplied by *the bridge*. The bridge is present *in* the sentence, allowing the sentence, overall, to say something *about* it. Generally, a display sentence features a constituent whose semantic role is to contribute itself. Searle notes that usually our subject matter is not within reach, hence we use a symbol to represent it. When the subject matter is in our vicinity, however, the possibility arises of *embedding* it within our discourse. In such

[17] His account of blindsight is essentially that the HOT represents as absent a visual percept that in fact obtains. Especially salient is his treatment of 'hallucination', in which 'noise', i.e. the absence of a relevant first-order state, is represented by the HOT as 'signal'—*producing every subjective appearance as of an existent first-order state*. Other worries about Lau's theory concern whether the probabilistic character of its HOT representations (compared with Rosenthal's decidedly *assertoric* HOTs) would produce experience, and whether the treatment of blindsight, which models it phenomenologically not as the absence of visual experience but as a visual experience as of nothing, is empirically accurate, not to mention conceptually coherent. As I explain below, it seems implausible that there can be positive experiences *as of nothing*.

[18] Kriegel (2009) derives the term 'display sentence' from Zemach (1985), who credits Searle (1969).

cases the term pertaining to the item does not *re*present it, but, being just the item itself, simply presents it (SEARLE, 1969).[19]

It is noteworthy that the bridge only acquires this presentational role when 'embedded' in the sentence. The bridge does not, standing alone, both exist as itself and present itself. That is a function of the context supplied by the semantic apparatus of the sentence wherein it features: within the sentence, it is *used* to present itself.

Kriegel's infectiously over-excited instinct is that display sentences hold the key to understanding consciousness: "There is something special and unusual going on […] which might help us feel a 'quantum leap' […] that might indeed be sufficient for [consciousness]." (KRIEGEL (2009, p. 164). His proposal is to model conscious states as *mental display sentences*. He considers that the best way to do this is via a sophisticated self-representational theory. Elsewhere I have argued that Kriegel's theory ultimately screens first-order states off from consciousness just as Rosenthal's theory does (COLEMAN, 2015). My own HOT theory amounts to a different way of constructing mental display sentences, which avoids the screening off of first-order sensory states. The guiding idea is that sensory states are constituents of the relevant consciousness-supporting states, and serve to display themselves within those constructions [see Zemach (1985, p. 196)].

This suggestion about experience is reminiscent of a popular theory of the structure of *phenomenal concepts*—those concepts we use to think directly about conscious states—*viz.*, that they 'quote' experiences (PAPINEAU (2002, 2007; BALOG, 2012). As Balog explains the notion of a quotational phenomenal concept:

The idea of an item partly constituting a representation that refers to that item is reminiscent of how linguistic quotation works. The referent of "__" is exemplified by whatever fills in the blank. In a quotation expression, a token of the referent is literally a constituent of the expression that refers to a type which it exemplifies and that expression has its reference (at least partly) in virtue of being so constituted. So, for example, ""dog"" refers to the word spelled d-o-g, a token of which is enclosed between the quotation marks… We can even imagine, perhaps just as a joke, placing something which is not a representation, e.g., a cat, between quotes and thus produce a representation that everyone can understand refers to the type cat. My proposal is that there

[19] Kriegel favours a *representational* reading of display, but I side with Searle, as explained below.

is a concept forming mechanism that operates on an experience and turns it into a phenomenal concept that refers to…a type of phenomenal experience that the token exemplifies (BALOG, 2012, p. 33).

My proposal is that the right higher-order analysis of *consciousness* sees a higher-order state 'quote' a first-order state, forming a larger composite structure wherein the first-order state is displayed. Its being embedded within the HO state, and thereby cognitively displayed, is what constitutes the subject's awareness of the state. In contrast to the Papineau/Balog model of phenomenal concepts, the quoted elements are not experiences—for my aim is to explain what turns first-order states *into* experiences. The suggestion is that it is an operation of *mental quotation*, hence this is the 'quotational higher-order thought' (QHOT) theory. When a first-order state is quoted in the requisite way, the result is a mental display sentence, which supports a conscious state, on the analysis.[20] Another difference is that QHOT theory deals in *tokens*: the question is what makes one conscious of a token first-order state. Balog's theory is concerned primarily with *types* of (experienced) states: typically we think about experience types, but we only experience sensory tokens. For Balog token experiences are recruited to represent experience types in thoughts about them. In QHOT theory mental state tokens are recruited for display in HO quotational structures that supply consciousness of said first-order content. The final, and crucial, difference is that QHOT theory's quotational structures are *nonrepresentational*: for there is no need to represent a token state which is actually present. This feature helps to remedy the screening-off problem provoked by the dual layer of contents on HOT theory (see §5).

The QHOTs that supply consciousness are very thin, best modelled as demonstrating 'frames', with a 'slot' for the first-order state. Rendered in language, a QHOT has a frame-like structure such as 'This state is present: "_____"', with the gap between the "_____" for the embedding of a first-order state. Let's insert our red London bus-ish percept to yield a complete instance of the schema: 'This state is present: "red London bus-ish visual quality"' The hypothesis is that a subject is conscious of this token red London bus-ish percept just in case she enters such a state.

[20] What is proposed is thus similar to Balog's 'joke': placing a non-representation within the QHOT's quotational structure.

5 Intimacy and mistargeting revisited

QHOT theory deals neatly with the mistargeting cases that, on a plausible reconstruction, largely motivate Rosenthal's decision to 'insulate' first-order states from consciousness, putting HOTs in charge of subjective mental appearances. But QHOT theory does this without screening off first-order states. Since first-order states get into the very fabric of the structures that constitute awareness of them, since they are displayed in those structures simply by being embedded, instead of being represented by an additional layer of content, this means that when we are conscious of a first-order content it is the very first-order state with this content of which we are aware.

The flip-side of this intimacy with our first-order states in consciousness is that consciousness cannot misconstrue its mental object: mistargeting is ruled out. First, misrepresentation cannot occur. A QHOT, as a bare quotational frame, determines no qualitative content of its own. Rather, all the qualitative content of a conscious state is supplied by the quoted first-order state. If I want quotationally to represent what Florence said yesterday in the heat of argument I can say 'She said: "Get out and never come back"'. A little closer to the QHOT model, I can play a tape-recording of what she said (was I sufficiently self-possessed to make one), saying 'She said this: *click*', i.e. playing the tape-recording at the relevant point. Closer still, I could summon Florence to repeat what she said, saying 'She said: "———"', and then letting her rip. I am not yet employing her *very utterance*, though, so we can imagine one further, somewhat outlandish, case. Had we a time machine, we could return to the instant Florence was about to shout at me, and I could say (observing my wretched past self): 'She said "———"', indicating the token utterance. Now, with this way of quoting what Florence said—*n.b.* it is no longer *represented*, but *exhibited*—I cannot stray from what she said, since her utterance is used by me to present itself. Similarly, it is impossible for a QHOT to misconstrue the first-order state of which it supplies awareness. For what supplies the qualitative content of the overall quotational conscious state is just that very first-order state *itself*, with *its* content. *This result is achieved by removing the dual layer of qualitative contents featured in HOT theory, and abandoning representation as the means by which the higher-order thought targets the first-order state, in favour of a quotation-style relation of part-whole constitution*.[21]

[21] See Coleman (2015) for more on how constitution can support non-representational intentionality. Cf. Kriegel's (2009) notion of constitution-based 'indirect representation'

What of targetlessness? A lone QHOT does not suffice for a conscious state. QHOTs are hypothesised to supply subjective awareness. However, if a QHOT has no first-order target to embed, it could at most arouse subjective awareness *as of nothing*, since it lacks qualitative content of its own. But subjective awareness as of a state of nothing at all is simply not subjective awareness—since subjective awareness must be (intentionally) of something. An experience literally of nothing is no experience. The thought 'This state is present: "*blank*"' in fact *fails* to be a genuine thought. Analogously, a linguistic quotational frame without any entrapped token *does not quote*, and fails to be a genuine sentence. QHOT theory, to repeat, construes consciousness as mental quotation. Thus the theory rules out targetlessness. [22]

It might appear that, like Block and Kriegel (and perhaps Neander too), I have unwarrantedly *presumed* that mental appearance and reality cannot diverge, and then sought to wield this presumption against Rosenthal and in favour of QHOT theory. But that is not so. QHOT theory is not motivated by the claim that mental appearance and reality cannot diverge, in the way that, on my diagnosis, the aforementioned critiques of HOT theory were. QHOT theory is motivated simply by the desirability of an account on which first-order contents are not screened off from consciousness by the contents of higher-order representations. The non-representational QHOT structure aims to deliver on that desideratum. It is then a *consequence* of the QHOT structure, and of the desirable conscious intimacy with first-order states that QHOTs deliver, that mistargeting cannot occur, *hence* that conscious appearances indeed coincide with conscious reality. But there is no presumption of that co-incidence as a starting point; it is strictly an upshot—albeit one that many theorists, myself included, are bound to find pleasing.

6 Conclusion

I have argued that a major historical line of objection to Rosenthal's HOT theory altogether fails, but that it nonetheless does reveal a reason to seek another account of consciousness—namely, the screening-off artifact we noted when it came to our consciousness of first-order mental states, as understood on HOT theory. I outlined QHOT theory as an alternative

[22] For more on how QHOT theory rules out misrepresentation and targetlessness see Coleman (2015).

HOT account that lacks this artifact.[23] QHOT theory achieves this result by removing the dual layer of contents that features in HOT theory, in favour of a non-representational model based on 'mental quotation'. As well as delivering a desirable intimacy with our first-order states in consciousness, QHOT theory has further pleasing consequences when it comes to the phenomena of misrepresentation and targetlessness that some theorists take to be a problem for Rosenthal's HOT theory: these phenomena turn out to be impossible. QHOT theory therefore has much to recommend it as an analysis of consciousness, even if more work admittedly remains to be done to explicate the notion of mental quotation and to identify possible mechanisms for its implementation in brains.

But, whatever the outcome of the internecine debate among HOT theories, what emerges clearly from the preceding discussion is that they possess great merits as a *genus* of theories of consciousness. As Rosenthal has rightly emphasised time and again, in order to find a suitable place for *consciousness* in the natural world, it is essential that we aim to devise an informative analysis of the concept and property, instead of resting content with an indissoluble mystery—as many philosophers nowadays do. This latter policy has disastrous consequences for the human project of self-understanding. For if consciousness is unanalysable, then it cannot be intelligibly related to the rest of the natural order.[24] The unacceptable alternatives we are then left with are that i). Humanity is *not* part of the natural order and our mental lives irrelevant to what goes on in the world, or ii). Humanity *is* part of the natural order, but not in respect of our conscious mental life, which turns out to be a species of epiphenomenon; an illusion that we must for some reason undergo.[25] As our leading attempt to analyse consciousness hitherto, it is therefore sincerely to be hoped that HOT theory, in one form or another, prevails!

COLEMAN, S. Os méritos das teorias do pensamento de ordem superior. *Trans/Form/Ação*, Marília, v. 41, p. 31-48, 2018. Edição Especial.

[23] N.b., as Coleman (2015) explains, screening-off of first-order states is a common feature of virtually all extant higher-order theories of consciousness.

[24] 'Russellian monist' versions of panpsychism might be thought to defy this blanket claim—see e.g. Alter and Coleman (2018), Goff (2017). But panpsychism faces strong objections—see Chalmers (2016), Coleman (2014, 2016).

[25] See Chalmers's (1996) zombies and the current trend to 'illusionism' about consciousness (DENNETT, 1991; FRANKISH, 2016) which this 'mysterian' approach to experience has produced.

Resumo: Durante muitos anos, e em muitas publicações, David Rosenthal desenvolveu, defendeu e aplicou sua justamente reconhecida teoria da consciência, intitulada *high-order thought* (HOT; pensamento de ordem superior). Neste artigo, explico a teoria e, em seguida, forneço uma breve história de uma objeção maior feita a ela. Sugiro que essa objeção é, em última análise, ineficaz, mas que por trás disso há uma razão para olhar além da teoria de Rosenthal, para outro tipo de teoria HOT. Então ofereço minha própria teoria HOT como uma alternativa adequada, antes de concluir, em uma seção final, a respeito de questões filosóficas aqui envolvidas.

Palavras-Chave: Consciência. Pensamento de ordem superior. Citação mental. Representação errônea. Familiaridade.

References

ALTER, T.; COLEMAN, S. Panpsychism and Russellian monism. *In*: SEAGER, W. (ed.). *The Routledge Handbook of Panpsychism*. London: Routledge, 2018.

ARMSTRONG, D. M. What is consciousness? *In*: HEIL, J. (ed.) *The nature of mind*. Ithaca, NY: Cornell University Press, 1981.

BALOG, K. Acquaintance and the mind-body problem. *In*: HILL, C.; GOZZANO, S. (ed.). *New perspectives on type identity:* the mental and the physical. Cambridge: Cambridge University Press, 2012. p. 16-42.

BLOCK, N. On a Confusion about a function of consciousness. *Behavioral and Brain Sciences*, v. 18, p. 227-287, 1995.

_____. Response to Rosenthal and Weisberg. *Analysis*, v. 7, n.13, p. 443-448, 2011.

BROWN, R. Review of Rocco J. Gennaro: the consciousness paradox: consciousness, concepts, and Higher-Order Thoughts. *Notre Dame Philosophical Reviews*, 2012. Availabe in: <http://ndpr.nd.edu/news/30848-theconsciousness-paradox-consciousness-concepts-and-higher-order-thoughts/1>. Access in: 10 ago. 2018.

BYRNE, A. Some like it HOT: consciousness and Higher-Order Thoughts. *Philosophical Studies*, v. 86, p. 103-129, 1997.

CASTON, V. Aristotle on consciousness. *Mind*, v. 111, n. 444, p. 751-815, 2002.

CHALMERS, D. J. *The conscious mind*: In Search of a Fundamental Theory. New York: Oxford University Press, 1996.

_____. The combination problem for panpsychism. *In*: BRUENTRUP, G.; JASKOLLA, L. (ed.). *Panpsychism*. New York: Oxford University Press, 2016.

COLEMAN, S. The real combination problem. *Erkenntnis*, v. 79, n. 1, p. 19-44, 2014.

_____. Quotational higher-order thought theory. *Philosophical Studies*, v. 172, n. 10, p. 2705-2733, 2015.

_____. Panpsychism and neutral monism: how to make up one's mind. *In*: BRUNTRUP, G.; JASKOLLA, L. (ed.). *Panpsychism*. New York: Oxford University Press, 2016.

DENNETT, D. *Consciousness explained*. Boston: Little, Brown, 1991.

FRANKISH, K. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, v. 23, n. 11/12, p. 11-39, 2016.

GOFF, P. *Consciousness and fundamental reality*. Oxford: Oxford University Press, 2017.

GOLDMAN, A. Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition*, v. 2, n.4, p. 364-382, 1993.

KRIEGEL, U. *Subjective consciousness*. New York: Oxford University Press, 2009.

KRIPKE, S. A. *Naming and necessity*. Cambridge, M.A.: Harvard University Press, 1972.

LAU, H. C. A Higher Order Bayesian decision theory of consciousness. *In*: BANERJEE, R.; CHAKRABARTI, B. K. (ed.). *Models of brain and mind:* physical, computational, and psychological approaches. Amsterdam: Elsevier, 2008.p. 35-48.

NEANDER, K. The division of phenomenal labour: a problem for representational theories of consciousness. *Philosophical Perspectives*, v. 12, p. 411-434, 1998.

PAPINEAU, D. *Thinking about consciousness*. Oxford: Oxford University Press, 2001.

_____. Phenomenal and perceptual concepts. *In*: ALTER, T.; WALTER, S. (ed.). *Phenomenal concepts and phenomenal knowledge:* new essays on consciousness and physicalism. Oxford: Oxford University Press, 2007. (Chap. 7).

ROSENTHAL, D. Varieties of higher-order theory. *In:* GENNARO, R. J. (ed.). *Higher-order theories of consciousness*. Amsterdam: John Benjamins, 2004. p. 17-44.

_____. *Consciousness and mind*. Oxford: Clarendon Press. 2005.

SEARLE, J. *Speech acts*. Cambridge: Cambridge University Press, 1969.

STUBENBERG, L. *Consciousness and qualia*. Amsterdam: John Benjamins, 1998.

ZEMACH, E. M. De se and Descartes: a new semantics for indexicals. *Nous*, v. 19, n. 2, p. 181-204, 1985.

Recebido: 15/11/2018
Aceito: 15/11/2018