Review of Philosophy and Psychology (2022) 13:295–309
https://doi.org/10.1007/s13164-022-00623-3

What is the Structure of Self-Consciousness and Conscious Mental States?

Rocco J. Gennaro[1]

Accepted: 27 January 2022 / Published online: 15 June 2022

Abstract
I believe that there is a ubiquitous pre-reflective self-awareness accompanying first-order conscious states. However, I do not think that such self-awareness is itself typically conscious. On my view, conscious self-awareness enters the picture during what is sometimes called "introspection" which is a more sophisticated form of self-consciousness. I argue that there is a very close connection between consciousness and self-consciousness and, more specifically, between the structure of all conscious states and self-consciousness partly based on the higher-order thought (HOT) theory of consciousness. A plausible notion of self-consciousness is, I think, simply having any degree of higher-order or meta-psychological thought. I argue that the connection between conscious states and self-consciousness is representational but also critically evaluate several different options. I then critique the alternative "acquaintance" theory of self-awareness and address a couple of recent criticisms of HOT theory. There is the potential danger of misrepresentation between self-awareness and conscious states which I also briefly address.

1 Introduction

Let me first summarize my views with respect to some of the guiding questions for this special issue on self-consciousness. I do believe that there is a ubiquitous pre-reflective self-awareness accompanying first-order conscious states. However, unlike others and for some of the reasons presented below, I do not think that such self-awareness is itself typically conscious. Thus, I reject the view that there is any conscious sense of "me-ness" or "mine-ness" which accompanies all conscious states. On my view, these enter the picture during what is sometimes called "introspection" (or "reflection") which is a more sophisticated form of self-consciousness. Still, I think that there is a very close connection between consciousness and

✉ Rocco J. Gennaro
rjgennaro@usi.edu

[1] Department of Political Science and Philosophy, University of Southern Indiana, Evansville, IN 47112, USA

self-consciousness and, more specifically, between the structure of conscious states and self-consciousness. A plausible notion of self-consciousness is, I think, simply having any degree of higher-order or meta-psychological thought. The connection between conscious states and self-consciousness is relational and representational as opposed to some kind of "acquaintance." In what follows, I first define and explain self-consciousness and my sympathy toward the so-called "higher-order thought" (HOT) theory of consciousness. Then I briefly discuss Sartre's somewhat similar view and then evaluate and differentiate among numerous representational accounts of self-awareness. I then critique an alternative "acquaintance" theory of self-awareness and address a couple of recent criticisms of HOT theory, including the issue of misrepresentation.

2 Self-Consciousness and HOT Theory

One question that should be answered by any theory of consciousness is: What makes a mental state a *conscious* mental state? A second question is: What is self-consciousness? On my own view, these two questions are also importantly related and having conscious mental states implies self-consciousness (Gennaro 1996, 2002, 2012). Still, there is significant disagreement about how best to understand this claim.

There is a long tradition that has attempted to understand consciousness as intimately tied to self-consciousness or some kind of higher-order awareness (e.g. Locke 1689/1975; Kant 1781/1965; see also Gennaro 1999 on Leibniz). One related way that I have sought to approach the issue is to defend a version of the higher-order thought (HOT) theory of consciousness which says that what makes a mental state M a conscious mental state is that there is a HOT to the effect that "I am in mental state M" (Rosenthal 1997, 2005; Gennaro 2012). The more general idea is that what makes a mental state conscious is that it is the object of a higher-order representation (HOR). A mental state M becomes conscious when there is a HOR of M. So, for example, my desire to drink a beer becomes conscious when I am (non-inferentially) "aware of" the desire. Intuitively, it seems that conscious states, as opposed to unconscious ones, are mental states that I am "aware of" in some sense. This is sometimes called the Transitivity Principle (TP):

(TP) A conscious state is a state whose subject is, in some way, aware of being in.

Conversely, the idea that I could be having a conscious state while totally *unaware* of being in that state seems odd or perhaps even contradictory. A mental state of which the subject is completely unaware is clearly an *un*conscious state. For example, I would not be aware of having a subliminal perception and thus it is an unconscious perception. For various reasons, HOT theorists prefer to characterize the HOR as a thought containing concepts, as opposed to a perception as in higher-order perception, or HOP, theory (Lycan 1996).

It is sometimes said that HOT theory results in circularity by defining consciousness in terms of HOTs. It also might seem that an infinite regress results because a conscious mental state must be accompanied by a HOT, which, in turn, must be accompanied by another HOT *ad infinitum*. However, the obvious and widely

accepted reply is that when a conscious mental state is a first-order world-directed state, the higher-order thought (HOT) is *not* itself conscious. When the HOT is itself conscious, there is a yet higher-order (or third-order) unconscious thought directed at the second-order state. In this case, we have *introspection* which involves a *conscious* HOT directed at a mental state. When one introspects, one's attention is directed back into one's mind or mental states. This seems to be a reasonable definition of introspection regardless of one's theory of consciousness. On the other hand, when one is having a first-order conscious state, one's attention is directed at the outer world (or perhaps at one's own body as in proprioception). As we shall see, this distinction is crucial to the discussion which follows.

HOT theory, like many contemporary theories of consciousness in analytic philosophy of mind, tends to focus on explaining "state" consciousness, that is, "what makes a mental *state* conscious?" We sometimes speak of an individual mental state, such as a pain or perception, as conscious. This contrasts with "creature" consciousness since we also often speak of organisms or creatures as conscious which is simply meant to refer to the fact that an organism is awake or aware of its surroundings, as opposed to sleeping or in a coma. Perhaps the most fundamental overall notion of 'conscious' among philosophers is captured by Thomas Nagel's famous "what it is like" sense (Nagel 1974). When I am in a conscious mental state, there is "something it is like" for me to be in that state from the first-person point of view. When I am, for example, smelling a rose or having a conscious visual experience, there is something it "seems" or "feels" like from my perspective.

To be more specific, I have argued for a version of HOT theory that I call the "wide intrinsicality view" (WIV). Unlike Rosenthal and others, I think that it is best to individuate conscious states "widely"; that is, to construe the HOT as an intrinsic part of an overall complex conscious state (Gennaro 2006, 2012 chapter four). So, according to the WIV, there are two parts of a single first-order conscious state with one part directed at ("aware of") the other. In short, we have a complex conscious mental state with an inner intrinsic relation between parts.

I think there are some advantages to the WIV over Rosenthal's version of HOT theory. For example: (1) The WIV can accommodate the intuitive belief that consciousness is an intrinsic property of mental states. From the first-person point of view, consciousness certainly seems to be an intrinsic feature of mental states (e.g. our visual perceptions). After all, consciousness does not seem to be an extrinsic property like "being to the left of." (2) The WIV also can explain the somewhat historically influential view that conscious mental states are, in at least some sense, reflexive or directed at themselves (Brentano 1874/1973).[1]

In any case, it also seems clear that there are various degrees of self-consciousness or self-awareness ranging from very minimal to very sophisticated. This is something

[1] Also: (3) If (or when) we better understand the neural correlates of consciousness (NCC's), it seems reasonable to think that they would be part of the conscious states themselves. (4) The WIV can better explain the essential interplay between cognition and concepts, on the one hand, and sensory states, on the other (Gennaro 2005, 2012). Nonetheless, much of what I have to say below applies both to the WIV and standard HOT theory.

which can also help to make sense of consciousness and self-consciousness in animals (Gennaro 2012, chapter eight). I think that a plausible notion of self-consciousness is simply having *any kind* of higher-order or meta-psychological thought. This can include any HOT, conscious or not, and even those containing such basic I-concepts as "bodily self-awareness"; that is, merely distinguishing one's own body from other things. So we are continuously "aware of" our mental states but this awareness is generally *not* a conscious awareness unless we are also attending to our mental states, in which case we have an unconscious third-order awareness of the second-order *introspective* awareness. According to HOT theory, of course, we can have a conscious state without introspecting it. It even seems useful to distinguish further between two kinds of introspection: momentary focused and deliberate (Gennaro 1996: 19). The former is a brief conscious focus on a mental state such as a pain or desire, whereas the latter involves the more sophisticated ability to consciously reason about our first-order mental states for a period of time. More sophisticated I-concepts than bodily self-awareness would include, for example, "I *qua* enduring thinking being" and "I *qua* experiencer of a current mental state."

Still, some might wonder why self-consciousness need not be consciousness of some thing. I offer two reasons here: (1) Few (if any) philosophers hold that self-consciousness is literally "consciousness of a self" especially since Hume observed that, even in introspection, we do not seem aware of an unchanging or underlying self but only a succession of mental states. (2) Although my definition of self-consciousness is weaker than some, other philosophers have proposed even weaker definitions. For example, Van Gulick (1988) urged that it is simply the possession of meta-psychological information. While I believe that this notion is too weak, my point here is only that my definition is not the weakest one in the literature.[2]

3 Sartre

To give just one brief example of a prominent philosopher who has argued for a related view to the above, Sartre makes numerous references to self-consciousness in much the same spirit. He says that "pre-reflective consciousness is self-consciousness" (Sartre 1956: 123) and that "this [non-positional] self-consciousness we ought to consider not as a new consciousness.... [it is] a quality of the positional consciousness" (Sartre 1956: 14). Indeed, Wider (1997) develops an interpretation of Sartre whereby "the most basic form of self-consciousness must be bodily awareness" (Wider 1997: 115).

I have argued that Sartre held something very close to the WIV in overall structure (Gennaro 2002). Non-positional (or pre-reflective) self-awareness is part of a conscious mental state. Sartre distinguishes between positional (or thetic) consciousness and non-positional (or non-thetic) consciousness. An act of consciousness is

[2] Some have attacked HOT theory and related views by arguing that so-called "depersonalization" psychopathological cases, such as thought insertion in schizophrenia and somatoparaphrenia, are problematic for HOT theory and related views. I respond to this line of argument in Gennaro 2015a, 2020, 2021.

"positional" or "thetic" when it asserts the existence of its object. Obviously related to the intentional nature of consciousness, the idea is that when ones conscious attention is focused on something else, one "posits" the existence of an intentional object. On the other hand, one merely has "non-positional" consciousness of "anything that falls within one's field of awareness but to which one is not now paying attention" (Wider 1997: 41). Every act of consciousness, Sartre eventually argues, has both a positional and non-positional aspect (but see Navas 2015 for a different interpretation).

Importantly, this self-awareness is not entirely separate from the mental state. When it comes to non-positional self-consciousness, it does not really posit an object, or at least not a *distinct* object. This is also why Sartre feels the need to explain that he uses the "of"[*de*] in parentheses merely out of "grammatical necessity" when speaking of non-positional (or pre-reflective) self-consciousness (of) self. But Sartre does also distinguish between pre-reflective (or unreflective) consciousness and reflective consciousness.

Still, Sartre interestingly explains that "there is no I on the unreflected level. When I run after a streetcar, when I look at the time, when I am absorbed in contemplating a portrait, there is no I. There is the consciousness of the streetcar-having-to-be-overtaken, etc., and non-positional consciousness of consciousness" (Sartre 1957: 48-49). Thus, I think that Sartre is saying that on the pre-reflective level there is no conscious apprehension of an I because my conscious attention is focused outside of me, as HOT theory also claims. The HOT theorist would say that there is no *conscious* I on the unreflected level when one has a first-order conscious state. There is, however, an unconscious or implicit I even on the unreflected level. Sounding again a bit like a HOT theorist, Sartre also recognized that "All reflecting consciousness is, indeed, in itself unreflected, and a new act of the third degree is necessary in order to posit it" (Sartre 1957: 45).

He was also aware of the potential for a problematic infinite regress: Either we stop at any one term of the series - the known, the knower known, the knower known by the knower, etc. In this case the totality of the phenomenon falls into the unknown; that is, we always bump up against a non-conscious reflection and a final term. Or else we affirm the necessity of an infinite regress…which is absurd...." (Sartre 1956: 12). Sartre is recognizing that when there is "reflecting" (i.e. introspective) consciousness, there must be "a new act of the third degree." This is reminiscent of the HOT theorist's contention that a third-order state is necessary for introspection. But, again, there is no infinite regress because a conscious state has no need for reflecting consciousness, that is, conscious mental states need not have a reflective (or introspective) state directed at them in order to be self-conscious.

4 Four Representational Theories

There are least four views in the relatively recent literature regarding the structure of conscious states and self-consciousness which invoke a representational relation between conscious states and self-awareness. For the sake of clarity and frequent comparisons, I will use the following notation from this point on:

- M = a world-directed (first-order) mental state
- M* = the meta-psychological or higher-order state directed at M
- M** = a third-order state directed at M*

I will use the acronym "**EHOT**" for Rosenthal's theory to emphasize that M* is entirely *extrinsic* to, or distinct from, M. We can call it "extrinsic HOT theory." So let us distinguish four positions with respect to first-order world-directed conscious states:

(**EHOT**) A mental state M of a subject S is conscious if and only if S has a *distinct or extrinsic* (unconscious) mental state M* (= a HOT) that is an appropriate representation of M.

(**WIV**) A mental state M of a subject S is conscious if and only if S has a suitable (unconscious) meta-psychological thought, M* (= MET), directed at M, such that both M and M* are *proper parts of* a complex conscious mental state, CMS.

(**PSR**) A mental state M of a subject S is conscious if and only if S has a mental state M* that is an appropriate representation of M, and M = M*.

(**SRT**) A mental state M of a subject S is conscious if and only if S has a (peripherally) conscious M* directed at M, such that both M and M* are *proper parts of* a complex conscious mental state, CMS.

All four views take seriously the intuitive notion that a conscious mental state M is a state that subject S is (noninferentially) aware that S is in (recall the TP mentioned earlier). As we have already seen with EHOT and the WIV, the differences lie in how each theory cashes out the expression "aware that one is in." **PSR**, or what might be called "pure self-referentialism" (Gennaro 2006) maintains that M* = M; that is, M is also literally directed back at itself, which seems closest to Brentano's (1874/1973) view and Kriegel's earlier view (2003). Brentano, using an example of hearing a sound or tone, explained that every mental act includes a consciousness of itself. Every conscious mental state has a double object, a primary and secondary object so, for example, you hear a sound *and* you hear the hearing of the sound. A conscious state is thus directed both outward toward an object and back at the entire state itself. Mental states, for Brentano, are essentially intentional. **SRT** (or self-representational theory) says that the M* in question is itself conscious but only peripherally so as opposed to the focal attentive nature of M (Kriegel 2009). In each case, some notion of self-reference, self-awareness, or reflexivity is involved and thus arguably some form of self-consciousness.

While both the WIV and EHOT theory answer the question "what makes a mental state a conscious mental state?" with something like "M becomes conscious when an appropriate unconscious HOT (M*) is directed at M," PSR can offer no such explanation. If we ask, "What *makes* M conscious?" for PSR, the response cannot be that M* is directed at M because M is supposed to be *identical with* M*. How can M* make M conscious or *explain* M's being conscious in any way if M* = M? Moreover, either M* is itself conscious or it is not, and then the familiar threat of regress (and even circularity) rears its ugly head. If M* is itself conscious, then what makes *it* conscious, and so on? Alternatively, if M* is not conscious, then the PSR defender would first have to acknowledge the existence of unconscious mental states

(which Brentano at least clearly does not), but even if so, how could M be conscious and M* be unconscious if M = M*?

A shift away from PSR and an emphasis on parts and wholes appears in Kriegel's (2009) book. He explicitly explains that there is really only an *indirect* self-representation to conscious states (2009: 215-226). Although I am sympathetic with this move to SRT since it is closer to the WIV, Kreigel's "indirect" self-representational theory maintains that the meta-psychological state in question [M*] is itself peripherally conscious and intrinsic to (or part of) the overall conscious state. However, I am not convinced that there is a ubiquitous conscious self-awareness, in the peripherally conscious sense, that accompanies each conscious first-order state (Gennaro 2008, 2012 chapter five). For example, it seems to me that our conscious attention is often so focused at the world and its objects that it seems unlikely that we are continuously *consciously* self-aware. It is also clear that such a view lacks plausibility *as compared to* the widely accepted and fairly obvious fact that there is always peripheral *outer-directed* conscious awareness, such as the awareness in one's peripheral visual field while working on one's computer.[3]

It is worth mentioning that Guillot (2017) makes a threefold distinction within what Kriegel sometimes calls "subjective character" (i.e. M*) as follows:

1. "**For-me-ness**" is when the "subjective character is a subject's characteristic awareness *of her experience*" (Guillot 2017: 31), that is, the object of self-awareness is the experience itself.
2. "**Me-ness**" is when "subjective character is a subject's awareness *of herself* as part of having the experience" (Guillot 2017: 31–32), that is, what makes an experience special for its subject is the fact that the subject is somehow aware of herself.
3. "**Mineness**" is when subjective character is a subject's "awareness of herself *as* having the experience" (Guillot 2017: 32); that is, awareness of herself as the owner of the experience.

On my view, however, assuming that each of the above three characterizations are meant to be itself conscious, none of them are present when one has a first-order conscious state since conscious attention is outer-directed. Thus, my position is what Farrell and McClelland (2017) call "absentism" which is that "inner awareness is never present in non-reflective experience and features only in introspective experience" (2017: 5; see also Howell and Thompson 2017). On the other hand, if for example, the phrase "the subject is somehow aware of herself" can be taken to include the presence of an I-concept in an *unconscious* HOT, then a HOT theorist could agree and so "me-ness" might be acceptable. It is important to note that we do often shift quickly back and forth between first-order and introspective states which may lead some to believe that one or more of the above is always present. It is also

[3] Yet another view would be Van Gulick's HOGS or "Higher-Order Global States" (see, for example, Van Gulick, 2004). I will not address HOGS in this paper because of space limitations and since I am somewhat less clear about how self-consciousness fits into it.

worth mentioning that those sympathetic with a Buddhist "no-self" view, or perhaps even Hume for that matter, may also reject at least one or two of the above three phenomena as present at all in first-order conscious states in the sense that there is no experiential phenomenology associated with *ownership* of experience (Chadha 2018).

5 The Acquaintance View

Some authors, however, have been entirely dissatisfied with thinking of the pre-reflective self-awareness in question as representational at all (such as Hellie 2007; Frank 2015; Preyer 2020). The idea is that it can be better construed as some kind of immediate "acquaintance" within our conscious states. I have previously criticized this view (Gennaro 2015b) but here I elaborate further and then defend the representational approach against additional criticisms from these authors.

There are, for example, numerous difficulties with the acquaintance approach and even just how to characterize it. For all of the talk of acquaintance being "immediate" and "nonrepresentational," it still remains the case that "acquaintance" is often described as a kind of *relation*, such as a relation between a subject and a conscious state or a relation between pre-reflective self-awareness and a conscious state. If it is not meant to be relational at all, then the onus is on the acquaintance theorist to explain it. To say that consciousness is "inseparable from" awareness of consciousness is not particularly helpful. I also agree here with Kriegel (2009: 106–113, 205–208) that the acquaintance strategy is at best trading one problem for an even deeper puzzle, namely, just how to understand the allegedly intimate and non-representational "awareness of" relation. I do not know what to make of this allegedly *sui generis* alternative. Indeed, acquaintance is often treated as unanalyzable and simple which makes it difficult to grasp how it could explain anything, let alone the structure and nature of conscious states or self-consciousness. How can there even be any "structure" at all if consciousness and awareness of consciousness is one and the same? There is really no *positive* description of acquaintance in the literature (see also Buras 2009; Zahavi 2007). It is even more difficult, if not impossible, to understand such intimate and reflexive "acquaintance relations" within the context of a neural realization.

Strawson (2015), who is also no friend of HOT theory, ultimately concedes that we should not give up the relationality claim in describing pre-reflective self-awareness or 'acquaintance.' He also doesn't see how genuine "reflexivity" or "self-reference" cannot involve genuine "relationality." So he is willing to allow that Sartre's 'of' in non-positional self-awareness is metaphysically accurate and not merely a way of speaking. He also rightly wonders how genuine reflexivity could not involve relationality. And even if one treats reflexivity as a kind of logical relation of "self-identity" (as perhaps Frank and others would), it would be too empty or trivial to explain anything specifically about acquaintance and conscious states – after all, everything is self-identical. Strawson also seems willing to agree that such "self-awareness" does not show up in the phenomenology of first-order conscious states.

On the other hand, HOT theory has some elements to work with, that is, representational relations between mental states or even state-parts. Still, it is fair to ask: So now how exactly does the relationship between two otherwise unconscious state-parts result in a conscious state? My own answer to this can be found in the way that concepts in HOTs are applied to first-order sensory representations (Gennaro 2012, chapter four). But anyone who doesn't wish to pursue a reductionist strategy will likely never think that consciousness can be explained in terms of "something else," similar to those who treat the so-called "hard problem" (Chalmers 1995) as a definitive reason to move toward dualism or to take a non-reductionist approach. It cannot be emphasized enough just how interconnected some views can be on both sides of this issue. In terms of the debate on pre-reflective self-awareness, for example, Brentano and Sartre did not even apparently think there were unconscious mental states and neither desired to offer a reductionist account of conscious states. It is no surprise that many in the more phenomenological tradition also tend to be anti-reductionist. Those who are anti-reductionist will then be more likely to embrace some notion of 'acquaintance' in accounting for conscious states and self-consciousness. On the other hand, those of us who wish to provide a reductionist account are naturally more open to representationalist approaches. We can also frame the issue as follows: If the acquaintance in question is itself conscious, then we are at serious risk of a regress similar to the way that PSR is susceptible to this worry, and it seems impossible to explain what makes an unconscious mental state a conscious one. If, on the other hand, the acquaintance theorist allows for "something else" to explain state consciousness, such as unconscious states or neural events, then any critique of EHOT or the WIV regarding their reductionist (and representational) strategy may apply equally to their theory as well.

HOT theory is also in a better position to explain what happens when there is a transition from a first-order conscious state to an introspective state, that is, an unconscious HOT becomes conscious. This transition can even perhaps be understood in evolutionary terms and, in particular, the evolution of the brain and corresponding mental capacity. We might naturally understand pre-reflective self-awareness as a stepping stone to reflective self-awareness. Even a non-HOT theorist might still agree with HOT theory as an account of introspection or reflection, namely, that it involves having a conscious thought about a mental state. It also seems uncontroversial to hold that when a mental state is unconscious, there is no HOT at all. But then there should be something in between those two cases when one has a first-order conscious state and HOT theory has a ready answer, namely, an unconscious HOT. It is unclear to me what an acquaintance theorist could say about such evolutionary development.

6 Frank's Critique of the Reflection Model of Self-Consciousness

Let us look further at Frank's criticisms of other theories (e.g. Frank 2015, 2019; see also Preyer 2020). Relying on "Fichte's Original Insight" and Heinrich's development thereof, he attacks what he calls the "reflection model" of self-consciousness. By "reflection model" he actually seems to have in mind any attempt to treat

self-consciousness, including pre-reflective self-awareness, as representational or intentional. Thus, he says that "self-consciousness cannot be explained as the result of a higher-order act, bending back upon a first-order one…" (Frank 2019: 36).

Frank explains that "Fichte held the opinion that….there is no inner pointing to a mental state or its owner in self-consciousness; self-consciousness is not the kind of consciousness in which objects are presented. Self-consciousness is radically *non-objectual*. It's not representational either, if representing is a two-place relation, whereas, according to Fichte, self-consciousness is precisely "immediate" in that subject and object of experience entirely, "seamlessly" coincide. There is no daylight between them" (Frank 2019: 45).

We have already seen the difficulties with how this notion of self-consciousness, or pre-reflective self-awareness, is not helpfully defined or understood. Frank then says that:

> "Rosenthal doesn't take umbrage at the consequence that [for HOT theory] there is no foreseeable end to the chain of reflections one superposed over the preceding one. The second act is on its part unconscious and becomes conscious by an act of third order. Such acts, Rosenthal comforts us, occur relatively rarely ("we would expect, [...], that the third-order thoughts that confer consciousness on such second-order thoughts would be relatively rare"). However, if a third-order thought is on its part unconscious, in order to get conscious it needs a thought of fourth order—*and that way the last thought would remain unconscious so that the entire chain collapses into the unconscious*" (Frank 2019: 49, emphasis added).

I find the above line of reasoning very puzzling. (1) Why should the "entire chain" collapse into the unconscious unless Frank is mistakenly supposing that it is the "top" mental state that is the one the subject is consciously experiencing at that time? To say that the entire chain "runs into the unconscious" simply points out that there is always an unconscious HOT at the top of any such chain. That is, during introspection, the third-order state is unconscious, but the second-order state is conscious because one is consciously attending to one's own mental state. On the other hand, during a typical first-order conscious state, such as the perception of a car, the second-order state is unconscious but the first-order state is consciously directed at the car. This is simply the way the theory works. One may disagree, of course, but it is begging the question to suppose that HOT theory cannot allow for unconscious HOTs or a straw man argument to think that the subject is consciously experiencing whatever is at the top of the chain. Rosenthal is simply pointing out hypothetically that "if" the third-order state were to be conscious, then a fourth-order HOT would need to be present and directed at it, and so on. But I am not aware of him or any HOT theorist supposing that we actually have any (or, at least, many) such HOTs – what kind of state would be a level up from introspection? It's not clear though perhaps I can have thoughts such as "I am thinking about my own thinking about self-consciousness." Still, Rosenthal points out that even introspective states do not occur often compared to first-order conscious states.

(2) Perhaps part of the problem can also be seen by recognizing that HOT theory is attempting to explain what Rosenthal (1993) calls *intransitive* consciousness as

opposed to *transitive* consciousness. Due to the lack of a direct object in the expression "x is conscious," this is usually referred to as *intransitive* consciousness, in contrast to *transitive* consciousness, where the locution "x is conscious of y" is used. So what makes the perception intransitively conscious? It is the unconscious transitive relation between its HOT and the perception. What would make an unconscious HOT (M*) itself intransitively conscious? The answer is the presence of an unconscious transitive relation between a third-order HOT (M**) and the target HOT. As we have seen, most contemporary theories of consciousness are aimed at explaining *state consciousness*, that is, what makes a mental state conscious. This is also the case for HOT theory in the sense that intransitive (state) consciousness is explained in terms of transitive consciousness.

Some have also accused HOT theorists, including myself, of running afoul of what they call the "*de se* constraint" on self-consciousness. For example, Borner et al. (2019) say that:

> "HOR theorists like Gennaro have tacked on conditions (e.g., not two distinct mental states but one state with two highly integrated parts, the one representing the other…)…But none of these moves address the initial point: HOR theory cannot deliver an explanation of the appearance of *self*-consciousness. Namely, what the HOR encounters in turning back (reflecting) upon a [first-order state] is not the second-order state or act *itself.* In so doing, Gennaro violates the *de se* constraint which requires that we call "self-conscious" only states in which the representing is directed at the represented *as at itself*" (Borner, Frank, and Williford: 5).

But, on my view, there is no phenomenologically conscious "appearance of self-consciousness" while having a typical outer-directed conscious first-order state. Thus, there is no violation of the *de se* constraint or, better, there is no need to agree that there is such a constraint at all in the way that some seems to have in mind. Like some other criticisms of HOT theory, the above line of argument seems to conflate pre-reflective and reflective (or introspective) self-consciousness. That is, when they talk about representing a mental state or oneself *as itself*, this seems to imply a more sophisticated form self-consciousness, namely, an introspective awareness of oneself as being in a given mental state. This clearly involves applying a rather sophisticated set of concepts. Indeed, the notion in question is sometimes explicitly described as involving the capacity to form I*-thoughts. The asterisk used here was introduced by Castañeda (1966) and can be found in the literature on *de se* propositional attitudes where an intentional state is directed *at oneself* (Perry 1979). It is not just that I can have thoughts or beliefs about myself but that *I myself* (or I*) realize that the thought or belief is about me. I may come to realize that the messy shopper accidentally dropping food in the aisles is me or suddenly realize that the person in the store mirror is me. So this is clearly a more sophisticated notion of self-consciousness where one is thinking about oneself *as an object*. As Baker (2013) explains, "I* is typically embedded in the "that-clause" of a complex first-person sentence with a psychological or linguistic main verb" (2013: 33), such as "I believe that I am short [where]…I am not only the thinker of the thought but I am also part of the object of my thought" (2013: 32-33). At minimum, we are at least disagreeing about the

appropriate definition and true nature of self-consciousness. But to treat the capacity for having I*-thoughts as definitive of self-consciousness seems to me to be overly restrictive in the sense that it refers to a rather advanced form of self-consciousness. This also makes me wonder how animals could even be conscious on this view.

When I *consciously* recognize that it is *me* who fits the description in question (e.g. I am the one who is messy or that is me in the mirror"), we have a case of introspection or reflection according to HOT theory, that is, the self-referential "I" becomes a conscious part of the state when the HOT itself is conscious via introspection. I may come to know or recognize that it is me who is dropping food in the supermarket aisle which I might not know by merely seeing the food. I can consciously experience that someone is doing something without consciously realizing that that someone is me. But, again, this sort of self-knowledge and self-awareness is a level up from instances of first-order conscious perceptions with unconscious HOTs. After all, merely having the conscious perceptions of the food does not require introspection.

7 Misrepresentation

One related difficulty for any representational account of pre-reflective self-awareness is what to say about the possibility of misrepresentation between M and M* which is frequently raised as an objection to HOT theory (Neander 1998; Levine 2001). How can one have a representational relation without the possibility of misrepresentation? Is the self-awareness of M* infallible with respect to M? On my view (the WIV), if the proper conceptual interconnectedness between M and M* is absent, then there will be no resulting conscious state. So, theoretically, if M* misrepresents M (or if there is no M at all), then what would otherwise be a complex conscious state does not exist and thus cannot be conscious. Misrepresentations cannot occur between M and M* (or MET) *and still result in a conscious state* (Gennaro 2012, chapter six).

On the neural level, much the same seems reasonable, since, for example, there may be some overlapping parts of feedforward and feedback loops that extend from M to M* or vice versa. It is well known that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons. The brain structures involved in these loops seem to resemble the structure of at least some form of HOT theory such that lower-order and higher-order states are combining to produce conscious states.

Once again, it is absolutely crucial not to conflate pre-reflective self-awareness with reflective self-awareness or, according to HOT theory, unconscious HOTs and conscious HOTs (i.e. introspection). This importantly applies to any discussion of knowledge claims about alleged infallibility. We must distinguish between claims of infallibility in outer-directed conscious states (i.e. between M and M*) with any sort of allegedly infallible *introspective* knowledge. In the WIV, it is possible to separate out the higher-order conscious state from its target mental state in cases of *introspection*. This is as it should be and does indeed allow for the possibility of error and misrepresentation. Thus, for example, I may

mistakenly consciously think that I am angry when I am really jealous, I may be wrong about why I did something, and so on. HOT theory and the WIV properly accommodates the widely held anti-Cartesian view that one can be mistaken about what mental state one is in, at least in the sense that when one *introspects* a mental state one may be mistaken about what state one is really in. There is more of an epistemic and metaphysical "gap" in the introspective case. However, this is very different from holding that the relationship between M and M* within a complex outer-directed conscious state is similarly fallible.

Further, even if one turned "knowledge by acquaintance" inward to one's own mental states, it wouldn't appear to be at the right level, i.e. at the pre-reflective level. Rather, it would seem to enter at the level of "introspection" or "reflection." Gertler (2011), for example, discusses this so-called "acquaintance theory of self-knowledge" as involving *introspection* of events in our minds. But she clearly recognizes that "acquaintance theorists are not generally committed to infallibility or omniscience," though introspection "can be more epistemically secure than other empirical judgments…The core of any acquaintance theory is the idea that some introspective knowledge involves acquaintance" (Gertler 2011: 96), as opposed, for example, to "inner-sense" accounts of introspective self-knowledge. HOT and other theorists might of course allow for some notion of "acquaintance" with regard to introspection, but just because the relation seems direct (e.g. without any conscious inference), it doesn't follow that it really is the way it appears or that it is infallible. Once again, the discussion at hand is clearly taking place at the introspective or reflective level, not at the pre-reflective level.

So, on my view, there is indeed a kind of infallibility *between M and M** but this is not a problem. The impossibility of error in this case is merely within the complex CMS and not some kind of certainty that holds between one's CMS and the outer object. When I have a conscious perception of a green car, I am indeed certain that I am having that perception, that is, I am in that state of mind. But this is much less controversial and certainly does not imply the problematic claim that I am certain that there really is a green car outside of me, as standard cases of hallucination and illusion are meant to show. If the normal causal sequence to having such a mental state is altered or disturbed, then misrepresentation and error can certainly creep in between my mind and outer reality. However, even in such cases, philosophers rarely, if ever, doubt that I am having the conscious state itself. I think the same should go for M and M* in EHOT theory though there is significant ongoing disagreement among HOT theorists on this matter (Weisberg 2011; Gottlieb 2020).

In closing, then, it is best to construe conscious states along the lines of HOT theory (or, more specifically, the WIV) which includes reference to oneself in the HOTs. Self-consciousness is having any kind of meta-psychological thought (including pre-reflective self-awareness) and thus not to be identified with introspection. Utilizing representational relations is also preferable to an acquaintance theory.

Data availability Not applicable

Code availability Not applicable

Declarations

Conflicts of interest/Competing interests Not applicable

References

Baker, Lynne Ruder. 2013. *Naturalism and the first-person perspective*. New York: Oxford University Press.

Borner, Marc, Manfred Frank, and Kenneth Williford. 2019. Introduction: pre-reflective self-consciousness and the *de se* constraint: The legacy of the Heidelberg School. *ProtoSociology* 36: 7–33.

Brentano, Franz. 1874/1973. *Psychology from an empirical standpoint*. New York: Humanities.

Buras, Todd. 2009. An argument against causal theories of mental content. *American Philosophical Quarterly* 46: 117–129.

Castañeda, Hector-Neri. 1966. "He": A study in the logic of self-consciousness. *Ratio* 8: 130–157.

Chadha, Monima. 2018. No-self and the phenomenology of ownership. *Australasian Journal of Philosophy* 96: 14–27.

Chalmers, David. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2: 200–219.

Farrell, Jonathan, and Tom McClelland. 2017. Editorial: consciousness and inner awareness. *Review of Philosophy and Psychology* 8: 1–22.

Frank, Manfred. 2015. Why should we think that self-consciousness is non-reflective? In *Pre-Reflective Consciousness: Sartre and Contemporary Philosophy of Mind*, ed. Sofia Miguens, Gerhard Preyer, and Clara Bravo Morando, 29–48. New York: Routledge Publishers.

Frank, Manfred. 2019. From "Fichte's original insight" to a moderate defense of self-representationalism. *ProtoSociology* 36: 36–78.

Gennaro, R. 1996. *Consciousness and self-consciousness: A defense of the higher order thought theory of consciousness*. Amsterdam and Philadelphia: John Benjamins Publishers.

Gennaro, R. 1999. Leibniz on consciousness and self-consciousness. In *New Essays on the Rationalists*, ed. R. Gennaro and C. Huenemann, 353–371. New York: Oxford University Press.

Gennaro, R. 2002. Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy* 32: 293–330.

Gennaro, R. 2005. The HOT theory of consciousness: Between a rock and a hard place? *Journal of Consciousness Studies* 12 (2): 3–21.

Gennaro, R. 2006. Between pure self-referentialism and the (extrinsic) HOT theory of consciousness. In *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford, 221–248. Cambridge, MA: MIT Press.

Gennaro, R. 2008. Representationalism, peripheral awareness, and the transparency of experience. *Philosophical Studies* 139: 39–56.

Gennaro, R. 2012. *The consciousness paradox: consciousness, concepts, and higher-order thoughts*. Cambridge, MA: The MIT Press.

Gennaro, R. 2015a. Somatoparaphrenia, anosognosia, and higher-order thoughts. In *Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness*, ed. R. Gennaro, 55–74. Cambridge, MA: The MIT Press.

Gennaro, R. 2015b. The 'of' of Intentionality and the 'of' of Acquaintance. In *Pre Reflective Consciousness: Sartre and Contemporary Philosophy of Mind*, eds. Sofia Miguens, Gerhard Preyer, and Clara Bravo Morando, 317-341. New York: Routledge Publishers.

Gennaro, R. 2020. Cotard syndrome, self-awareness, and I-concepts. Philosophy and the Mind Sciences (Special Issue: Radical Disruptions of Self-Consciousness), 1: 1-20. Available at: https://doi.org/10.33735/phimisci.2020.I.41

Gennaro, R. 2021. Inserted thoughts and the higher-order thought theory of consciousness. In *Psychiatry and Neurosciences Update*: Vol 4, eds. Pascual Angel Gargiulo and Humberto Luis Mesones-Arroyo, 61-71. Dordrecht: Springer.

Gertler, Brie. 2011. *Self-Knowledge*. London: Routledge Publishers.
Gottlieb, J. 2020. On ambitious higher-order theories of consciousness. *Philosophical Psychology* 33: 421–441.
Guillot, Marie. 2017. I me mine: on a confusion concerning the subjective character of experience. *Review of Philosophy and Psychology* 8: 23–53.
Hellie, Benj. 2007. Higher-order intentionality and higher-order acquaintance. *Philosophical Studies* 134: 289–324.
Howell, Robert, and Brad Thompson. 2017. Phenomenally mine: In search of the subjective character of consciousness. *Review of Philosophy and Psychology* 8: 103–127.
Kant, Immanuel. 1781/1965. *Critique of pure reason.* Translated by N. Kemp Smith. New York: MacMillan.
Kriegel, Uriah. 2003. Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy* 33: 103–132.
Kriegel, Uriah. 2009. *Subjective consciousness*. New York: Oxford University Press.
Levine, Joseph. 2001. *Purple haze: The puzzle of conscious experience*. Cambridge: MIT Press.
Locke, John. 1689/1975. *An essay concerning human understanding.* P. Nidditch ed. Oxford: Clarendon.
Lycan, William. 1996. *Consciousness and experience*. Cambridge: MIT Press.
Nagel, Thomas. 1974. What is it like to be a bat? *Philosophical Review* 83: 435–456.
Navas, Daniel. 2015. Does consciousness necessitate self-awareness?: Consciousness and self-awareness in Sartre's *The Transcendence of the Ego*. In *Pre-Reflective Consciousness: Sartre and Contemporary Philosophy of Mind*, ed. Sofia Miguens, Gerhard Preyer, and Clara Bravo Morando, 225–244. New York: Routledge Publishers.
Neander, Karen. 1998. The division of phenomenal labor: a problem for representational theories of consciousness. *Philosophical Perspectives* 12: 411–434.
Perry, John. 1979. The problem of the essential indexical. *Nous* 13: 3–21.
Preyer, Gerhard. 2020. Concepts of consciousness and representation. Merits and critiques of higher and same order monitoring accounts in the theories of the mental. *Studia Historii Filozofii* 1: 95–120.
Rosenthal, David. 1993. State consciousness and transitive consciousness. *Consciousness and Cognition* 2: 355–363.
Rosenthal, David. 1997. A theory of consciousness. In *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Guven Güzeldere, 729–753. Cambridge: MIT Press.
Rosenthal, David. 2005. *Consciousness and mind*. New York: Oxford University Press.
Sartre, Jean-Paul. 1956. *Being and nothingness*. New York: Philosophical Library.
Sartre, Jean-Paul. 1957. *The transcendence of the ego, trans. Forrest Williams and Robert Kirkpatrick*. New York: Hill and Wang.
Strawson, Galen. 2015. Self-Intimation. *Phenomenology and the Cognitive Sciences* 14: 1–31.
Van Gulick, Robert. 1988. A functionalist plea for self-consciousness. *The Philosophical Review* 97: 149–181.
Van Gulick, Robert. 2004. Higher-order global states (HOGS): an alternative higher-order model of consciousness. In *Higher-order theories of consciousness: an anthology*, ed. Rocco Gennaro, 67–92. Amsterdam and Philadelphia: John Benjamins Publishers.
Weisberg, Josh. 2011. Misrepresenting consciousness. *Philosophical Studies* 154: 409–433.
Wider, Kathleen. 1997. *The bodily nature of consciousness: Sartre and contemporary philosophy of mind*. Ithaca: Cornell University Press.
Zahavi, Dan. 2007. The Heidelberg School and the limits of reflection. In *Consciousness: From perception to reflection in the history of philosophy*, ed. S. Heinämaa, V. Lähteenmäki, and P. Remes, 267–285. Dordrecht: Springer.