Two Kinds of Introspection

Anna Giustina & Uriah Kriegel

Forthcoming in J. Weisberg (ed.),

Qualitative Consciousness: Themes from the Philosophy of David Rosenthal, CUP.

Introduction/Abstract

One of David Rosenthal's many important contributions to the philosophy of mind was his clear and unshirking account of introspection. Here we argue that while there is a kind of introspection (we call it "reflective introspection") that Rosenthal's account may be structurally fit to accommodate, there is also a second kind ("primitive introspection") that his account cannot recover. We introduce Rosenthal's account of introspection in §1, present the case for the psychological reality of primitive introspection in §2, and argue that Rosenthal's account lacks the resources to accommodate it in §3.

1. Introspection as Third-Order Thought: Rosenthal's Account

1985 was a long time ago. Reagan was sworn in for a second term after winning 97.6% of the electoral vote. The internet's domain name system was established. In early July, *Back to the Future* hit the theaters across the US. A few days later, on 8 July 1985, an article landed in the offices of *Philosophical Studies* in Tucson, Arizona, titled 'Two Concepts of Consciousness.' The plucky young author, one David M. Rosenthal, opened it thus:

> No mental phenomenon is more central than consciousness to an adequate understanding of the mind. Nor does any mental phenomenon seem more stubbornly to resist theoretical treatment. (Rosenthal 1986: 329)

He went on to develop what is arguably the first clear and precise account of the nature of consciousness and introspection in analytic philosophy of mind.

Rosenthal's "higher-order thought" theory of consciousness proceeded to exert immense influence on, and attract constant discussion in, contemporary philosophy of consciousness. Here we want to focus on Rosenthal's theory of introspection, which is no less clear and precise but has received less attention. Rosenthal's theory of introspection does flow seamlessly from his theory of consciousness, though, so we start our discussion with the latter.

Rosenthal's theory of consciousness can be seen as a combination of two parts: (i) a fundamental principle, intuitive and pre-theoretically compelling, that operates as a kind of datum, and (ii) a theoretical edifice erected around this datum and designed to do justice to it while respecting a diverse collection of desiderata and (both a priori and a posteriori) plausibility considerations.

The fundamental principle is Rosenthal's so-called transitivity principle, which was later to become a central decision point for all major philosophical theories of consciousness. The principle is simple: Conscious states are states we are aware of. Unconscious mental states may occur in us without our having any inkling that they do, but a conscious state is different – we are "to some degree aware of being in it" (Rosenthal 1986: 334). The principle is straightforward, but it turns out that, combined with a series of independently plausible considerations, it generates a rather comprehensive and textured characterization of the nature of consciousness. Suppose S has a conscious desire for pistachio gelato while S* has an *unconscious* desire for pistachio gelato. What does the psychological difference between S and S* amount to exactly? Bracketing some subtleties, Rosenthal's answer may be summarized as follows: S's desire is, whereas S*'s is not, the object of a non-inferential higher-order thought. In other words, what makes a mental state conscious is that its subject also has another mental state that represents it, where that other mental state is a non-inferential higher-order thought.

Importantly, in Rosenthal's theory this higher-order thought does not *do* anything to the lower-order state in order to make it conscious. It does not bring about any *intrinsic change* in that state that *renders* that state conscious. Rather, the higher-order thought makes the lower-order state conscious simply by *being there*. It is in this sense that consciousness is, in Rosenthal's theory, a *relational* rather than intrinsic property of conscious states (1986: 354). A mental state becoming conscious is, in this framework, a *Cambridge change* (the kind of change one undergoes, e.g., when one becomes an uncle). We can have two intrinsically indistinguishable mental states only one of which is conscious, namely, if only one of them happens to occur at the same time as some suitable higher-order thought. This point is quite crucial to Rosenthal's theory's reductive ambitions, its attempt to account for the nature of consciousness in terms of the

coming-together of elements none of which is conscious on its own. If the higher-order thought needed to in any way *change* the lower-order state to make it conscious, then the lower-order state's consciousness would consist in whatever intrinsic modification would be thereby effected, and the nature of consciousness would be identical with the nature of this intrinsic modification. The higher-order thought's role with respect to consciousness would only be to *causal*, not *constitutive*—that is, it would concern what makes consciousness occur, not what consciousness *is*. But it is only by giving the higher-order thought a *constitutive* role in consciousness, offering an account of what consciousness *is*, that Rosenthal can make good on his promise to "get underneath" consciousness and reductively account for it in terms of the coming-together of non-conscious elements.

How, then, does Rosenthal reach his reductive theory of consciousness, whereby a mental state is conscious just when it co-occurs with a non-inferential higher-order thought about it, from the fundamental principle that conscious states are states we are aware of? The answer is that he obtains it by conjoining the principle with three independent considerations, which we will now briefly expound.

1. Assuming, then, that when S has a conscious desire for pistachio gelato, the desire is conscious because S is aware of it, why should we construe S's awareness of her desire as a *thought*? Well, it is surely not a *desire*, since unfortunately we are often in conscious states we wish we were not in. Thus, S may be on a diet and may wish she did not want gelato so often. But nor is S's awareness of her desire a *perception* (as "higher-order perception theories" suggest), since perception is produced by sense organs, and there is no organ of inner awareness. We can see this by considering that the external senses are associated with distinctive qualities (visual, auditory, etc.), whereas our awareness of our conscious states is associated with no distinctive qualities (Rosenthal 1990: 740). (Could this awareness still be "quasi-perceptual"? It depends on what we mean by this, but according to Rosenthal, as long as there are no sensory qualities associated with awareness of conscious states, any analogy to perception would be "idle" – *Ibid.*) If S's awareness of her desire for gelato is neither a desire nor a (quasi-)perception, then plausibly it is just a thought.

2. Why is S's thought about her desire *higher-order*? The alternative to a conscious state being the object of a higher-order thought is that it be its own object (as "self-representational" theories claim). According to this alternative, S's desire and her thought about her desire are one and the same mental state. But, claims Rosenthal (1990: 746-7), mental states individuate by content and

attitude, and a desire with the content <I eat pistachio gelato> differs both in content and in attitude from a thought with the content <I desire that I eat pistachio gelato>. Since they differ both in content and in attitude, concludes Rosenthal, they must be distinct mental states.

3. Why must S's higher-order thought be *non-inferential*? To be clear, here we reserve the term "inference" to personal-level acts that a subject consciously performs; there are certainly inference-like unconscious processes, we just do not call them "inference." In these terms, then, the question is why it is indispensable that S's thought that she desires pistachio gelato not be formed on the basis of conscious inference? The basic reason is that allowing the higher-order thought to be formed by conscious inference would produce an extensionally inadequate theory. One can perfectly well infer from one's behavior, or from another's testimony, that one is in some *unconscious* mental state (Rosenthal 1990: 737). But there is also a deep reason for this extensional inadequacy: the kind of awareness of our mental states that makes them conscious is an *immediate* awareness – so it cannot be mediated by conscious inference (Rosenthal 1986: 335-6, 1993).

There is a fourth characteristic Rosenthal ascribed to consciousness-making higher-order thoughts: although they confer consciousness on their objects, typically they themselves are *unconscious*. The reason for this is straightforward: if the higher-order thought about the desire for gelato were itself conscious, then by the theory's own lights, it would have to be the object of a *third*-order thought, and we would soon be off on a vicious regress.

Nonetheless, claimed Rosenthal, it is psychologically possible for us to have conscious second-order thoughts accompanied (and made conscious) by unconscious third-order thoughts about them. In fact, this is exactly what *introspecting* amounts to (Rosenthal 1986: 353-4). Introspection is often construed as a matter of entering a second-order state that represents one's current conscious experience. But within Rosenthal's theory of consciousness, the subject is in this kind of second-order state *whenever* she is in a conscious state, regardless of whether she is introspecting her conscious state or not. Accordingly, for Rosenthal the transition from being in a non-introspected conscious state to introspecting that state is rather a matter of transitioning from being in an *unconscious* second-order thought to a *conscious* one. Given Rosenthal's theory of what makes a mental state conscious, this means that the introspecting subject is the subject who (non-inferentially) enters a *third*-order state, namely, a(n unconscious) thought about her second-order thought. Thus to be introspectively aware of desiring gelato is to be in three simultaneous but distinct mental states: a conscious first-order desire with a content roughly like

<I eat gelato>, a conscious second-order thought with a content roughly like <I desire that I eat gelato>, and an unconscious third-order thought with a content roughly like <I think that I desire that I eat gelato>. Call this the "Third-Order Thought model of introspection."

Here we can see how Rosenthal's theory of introspection flows directly from his theory of consciousness. If the basic "datum" for the theory of consciousness is that conscious states are states we are aware of, the corresponding datum for the theory of introspection is that introspected conscious states are states we are *consciously* aware of: we have not just any non-inferential higher-order thought about them, but a *conscious* one. By then reapplying the account of what makes a first-order state conscious to the case of second-order states, we obtain a theory of introspection.

Naturally, here too the third-order thought does not *do* anything to the second-order thought it is about. It does not render the second-order state conscious by bringing about any intrinsic change in it. Rather, the third-order thought makes it the case that the second-order thought is conscious, hence that the subject is in an introspective state, merely by *being there*. Introspection is relational, then, in the same way consciousness is. This is only to be expected, since in this framework introspection is a matter of transitioning from an unconscious to a conscious second-order thought.

(Are there cases involving a *conscious* third-order thought, made conscious by an unconscious *fourth*-order thought? It depends on what the contingent, empirical laws of psychology permit. The higher-order theory need not take a stand on this. *If* this kind of "second-order introspection" is part of our psychological repertoire, *then* the theory would model it in terms of a hierarchy of four simultaneous mental states. But if second-order introspection is not in the cards for us, then there is nothing for the theory to model.)

Note that although a subject being in an introspective state involves, in Rosenthal's theory, a three-tier structure with three distinct states, the introspective state proper is only the second-order state. The first-order state is rather the introspect*ed* state, while the third-order state is the state that makes the introspective state conscious. Accordingly, it is the properties of the second-order state in this three-tier structure that are the properties of the introspective state in Rosenthal's theory.

The property we will focus on in what follows is that of *being a thought*. It is a feature of Rosenthal's theory that all introspective states are thoughts. For an introspective state just is the state that renders the first-order state conscious when it itself becomes conscious, and the states that render first-order states

conscious are always, in Rosenthal's theory, *thoughts*. What does it mean to say that they are thoughts? It means, at a minimum, that they have the kind of content that thoughts have and take the kind of attitude toward that content that thoughts take. The content that thoughts have is conceptual content with propositional structure, that is, a proposition the constituents of which are concepts. The attitude thoughts take toward this sort of content is a belief-like attitude with a mind-to-world direction of fit, what Rosenthal calls "assertoric force" (Rosenthal 1986: 346-7, 1990: 742.) Thus, S's introspective state is a state that takes an assertoric attitude toward the content <I desire that I eat gelato>, which content deploys the concepts of self, desire, eating, and gelato and has the propositional structure *a*R*b* (where *a* is the concept of self, R is the concept of desire, and *b* is a structured composite of three concepts: self, eating, and gelato).

The reason we focus on the thought-y nature of the second-order state in Rosenthal's theory of consciousness and introspection is that in §2 we will argue for the psychological reality of an introspective phenomenon we call "primitive introspection" (Giustina 2018, 2019), which is *not* a thought: its content is *not* the kind of content that thoughts have. This will create a prima facie difficulty for Rosenthal's theory of introspection. The purpose of §3 will then be to bring this prima facie difficulty nearer the status of an ultima facie argument against the theory.

2. Reflective Introspection and Primitive Introspection

Introspection is commonly characterized as a distinctively first-personal method of getting knowledge of one's current phenomenally conscious states. By "distinctively first-personal" we mean that this method is available to the person whose phenomenal state is introspected in a way it is not to other persons. By "current" states, we mean states roughly simultaneous with the introspecting. As we have seen, Rosenthal's account of introspecting takes it to be a thought, and therefore to display at least the following two features: (1) it is *conceptual*, which at a minimum means it involves the deployment of a mental representation that enables the subject to (a) distinguish Cs from non-Cs and (b) recognize token Cs as instances of the *type C*; (2) it is *propositional*, i.e. has a predicatively structured content which is made up of concepts. It follows that all introspection is *classificatory*: it involves classifying or recognizing what is introspected as an instance of a certain experience type.

Of course, we do not deny the existence of an introspective phenomenon answering roughly to this characterization. On our view, however, there is also *another* important introspective phenomenon. Beside this classificatory, thought-like kind of introspection, which may be called "reflective introspection," there is a kind of introspection that does *not* have the form of a thought; we call this "primitive introspection." Primitive introspection is a kind of introspection of phenomenal states that does *not* involve classifying the introspected state as an instance of any state type.[1]

Consider Sara's gustatory experience when she tastes pistachio gelato. Sara has had pistachio gelato several times before, and thus knows what pistachio gelato tastes like and can recognize a pistachio-gelato taste experience when she has one. Accordingly, when she introspects her taste experience, Sara introspects it *reflectively*: she immediately classifies it as pistachio-gelato taste experience and thereby forms an introspective thought with a content roughly like <this is a pistachio-gelato taste experience> (or perhaps <I am having a pistachio-gelato taste experience>). Unlike Sara, however, Sam has never tasted pistachio gelato. When he tastes it for the first time, he is able of course to introspectively attend to his taste experience; but he is *not* able to classify it as pistachio-gelato taste experience, because he cannot recognize it as such. Now, it may be objected, correctly, that there are *some* introspective thoughts Sam can form about his experience—say, that it is a gelato-taste experience, or at the very least that it is a *taste* experience. But consider Sacha, whose rare condition has prevented him from having any taste experience until now. When a cure is mercifully found, Sacha's first taste experience is pistachio gelato. Here there is no already-encountered experience type such that Sacha could classify his experience as an instance of it. Nevertheless, Sacha can still introspect his taste experience: he can become introspectively aware of his experience even though he cannot form *any* introspective thought about it. Sacha's introspective state is what we call a state of *primitive introspection*.

Sacha's case may appear a bit far-fetched, but there are, arguably, at least three kinds of everyday-life case featuring primitive introspection. *First*, average humans do sometimes have categorically new experiences, which therefore they cannot quite classify: when one has an orgasm for the first time, say, one cannot recognize or classify it as an instance of a previously encountered experience type. Nevertheless, one can certainly attend to the experience and thereby

[1] The scope of such phenomenal-state introspection partly depends on what conscious states have phenomenology. Whereas for some conscious states (e.g. perceptual, algedonic, bodily) there is virtually unanimous agreement that they do have phenomenology, others (e.g. cognitive states) are object of controversy. We remain neutral on this and simply assume that whatever states have phenomenology, they are potential targets of primitive introspection.

introspect it. *Second*, there are cases where, although one *could* classify the experience, one chooses not to do so, as when one just wants to contemplate the phenomenology of one's experience without attaching any judgment to it (some meditation practices claim to seek just such a contemplative introspective state). *Third*, even cases in which one *does* classify the introspected experience may nonetheless feature primitive introspection: they may be simply cases in which primitive introspection and reflective introspection co-occur. When we attend to a visual experience of an intricate mandala, the phenomenology of our experience is extremely rich and its fineness of grain is difficult (if not impossible) to capture by our classificatory abilities. Arguably, whatever introspective judgment we form about our visual experience, we *also* have a state of primitive introspection that captures the details of the phenomenology that go beyond what shows up in our introspective judgment. (For more on this, see Giustina 2018 Ch.1.)

There are also theoretical considerations that support the psychological reality of primitive introspection. If we did not admit the existence of a non-conceptual, non-classificatory kind of introspection, the acquisition of most phenomenal concepts (concepts associated with the phenomenology of experiences and deployed in introspective thoughts) would be mysterious. For how does one come to possess a phenomenal concept C? There seem to be only three options here: (i) C is acquired by introspection, (ii) C is acquired, but not through introspection, (iii) C is innate rather than acquired. Now, although options (ii) and (iii) may be viable for *some* phenomenal concepts, they surely are not viable for *all* or even *most* of them.

Consider option (iii). Although many respectable theories posit *a few* innate concepts, the view that *all* or *most* phenomenal concepts are innate is hard to believe. For one thing, it entails that newborns possess mental representations that enable them to (a) discriminate a great number of experiences and (b) recognize each of them as an instance of a certain experience type. This seems quite implausible. Moreover, it would have the counterintuitive consequence that a newborn possesses, say, the phenomenal concept TRUMPET SOUND even if she has never heard any trumpet (indeed, even if she has never heard *anything*), or that she possesses the phenomenal concept PISTACHIO-GELATO TASTE before having tasted anything at all. In the same vein, a congenitally blind person would possess phenomenal concepts of color experiences, a congenitally deaf person would possess phenomenal concepts of sound experiences, and so on. All this is very hard to believe.

As for option (ii), it is highly plausible that *some* phenomenal concepts are acquired otherwise than by introspection, notably through composition of other

concepts. The phenomenal concept MILD PAIN, for instance, may well be built up from MILD and PAIN. Obviously, however, *not all* phenomenal concepts can be formed compositionally, on pain of infinite regress. Some phenomenal concepts must be atomic or simple. Of these, some may still be non-introspectively acquired – perhaps through *perception*. Some philosophers hold that phenomenal concepts of perceptual experiences (PHENOMENAL-RED, PHENOMENAL-SQUARE, etc.) are acquired by simply attending to the external object perceived, rather than by introspecting the relevant perceptual experience. However, for other kinds of experience, notably emotional and mood experiences, this is far less plausible. Even if fear involves representation of something as dangerous and fury involves representation of something as offensive, it does not seem possible to acquire the concepts of fear and fury merely by attending to dangerous and offensive things. Plausibly, the concepts we acquire by attending to dangerous and offensive things are precisely the concepts DANGEROUS and OFFENSIVE; but the concept FEAR is not identical to the concept DANGEROUS and the concept FURY is not the same as the concept OFFENSIVE. The concepts DANGEROUS and OFFENSIVE can be employed by someone completely incapable of experiencing fear and fury, but arguably such a person could not possess the phenomenal concepts FEAR and FURY. Accordingly, acquiring some phenomenal concepts clearly requires more than attending to the objects in one's environment and their features: one needs to *introspectively attend to* one's experience of these objects.

We conclude that (i) must be true of *some* phenomenal concepts: some such concepts must be acquired through introspection. But if all introspection were conceptual, this would be impossible, on pain of circularity. Obviously, if a subject S's having an introspective state φ depends on S's deploying a previously possessed phenomenal concept C, C must be possessed by S *prior to* having φ and cannot itself be acquired by S *through* being in φ. (For a fuller development of this argument from concept acquisition, see Giustina 2019.)

All these considerations, we suggest, speak up for the psychological reality of primitive introspection. If primitive introspection is real, now, then any adequate theory of introspection should be at least consistent with, and ideally account for, the existence of this kind of introspective state. In the next section, we argue that Rosenthal's Third-Order Thought model of introspection does not have the resources to do so. Although the model is structurally fit to account for reflective introspection, it does not accommodate primitive introspection.

3. Can the Third-Order Thought Model Accommodate Primitive Introspection?

The Third-Order Thought (TOT) model of introspection entails that all introspective states are thoughts. However, primitive-introspective states are *not* thoughts. Therefore, the TOT model seems unfit to capture primitive introspection. Our basic argument is this, then:

(P1) There is a non-classificatory form of introspection.

(P2) If all introspective states were second-order thoughts, then there would not be a non-classificatory form of introspection.

(C) Not all introspective states are second-order thoughts.

Support for (P1) comes from the arguments we outlined in §2. (P2) is a direct consequence of Rosenthal's theory of introspection, as seen in §1. In what follows, we consider potential objections to (P1) and (P2) and argue that they can be rebutted.

3.1. Defending (P1)

Rosenthal would likely refuse to accept the very existence of primitive introspection. We offered two main motivations for primitive introspection. One was reflection on everyday cases, of at least three kinds: where the subjects does not know *how* to classify (first-orgasm case), where the subject does not *want* to classify (meditation case), and where despite classifying her experience the subject is also primitive-introspectively aware of certain aspects of the experience that she *cannot* classify (mandala case). The other motivation was the argument from phenomenal-concept acquisition: the idea that primitive introspection is the best explanation of the acquisition of at least some phenomenal concepts.

As regards the everyday cases, Rosenthal might try to accommodate them in terms of *comparative* HOTs. When, for instance, we introspect a phenomenally red experience of a (phenomenal) shade too fine-grained for us to have a concept for, Rosenthal (2005: 188-9) argues that we can still apply to it *comparative* concepts, such as SLIGHTLY BRIGHTER and SLIGHTLY BLUER. Perhaps the strategy can be extended to our cases? The strategy would not apply to cases where the subject does not *want* to classify her experience, but perhaps Rosenthal would insist – much to Zen monks' chagrin – that this is a desire we cannot satisfy.

The problem with the appeal to comparative concepts is that it does not speak to cases where we claim primitive introspection *accompanies* reflective introspection. Even if, upon introspecting a mandala experience, we apply to it the comparative concept MUCH MORE INTRICATE, or indeed the noncomparative concept MANDALA EXPERIENCE, there is also something further we are aware of in our experience, namely the *specific mandala phenomenology* of our experience. We are aware of our experience not just as being more intricate than some other experiences but as being this particular intricate shape phenomenology, a shape phenomenology for which we have a concept. In truth, this seems to apply to the case of phenomenally red experience as well: the (concept-outstripping) specific shade of red shows up in our introspective awareness too, not just its slightly-brighter-ness and its slightly-bluer-ness!

As concerns the concept-acquisition argument, one possible objection might be this. In defending the claim that some phenomenal concepts must be acquired through introspection, we have considered three alternatives: (a) that phenomenal concepts are innate, (b) that they are acquired through perception, and (c) that they are acquired through composition. However, we considered them only *in isolation*. We have not considered the possibility (d) that all phenomenal concepts are either innate or acquired through a combination of perception and composition. In particular, Rosenthal might propose the following view: there is only *one* innate phenomenal concept, namely the concept EXPERIENCE, and the other phenomenal concepts are acquired by composition of this concept and some *perceptual* concepts. The phenomenal concept FEAR, for instance, might be acquired by combining EXPERIENCE with DANGEROUS (on the assumption that experiencing fear implies perception of danger)—plus, perhaps, some proprioceptive perceptual concepts of the kind of bodily changes that typically accompany fear experiences. Similarly, the phenomenal concept FURY might be acquired by combining EXPERIENCE with OFFENSIVE (plus fury-related bodily-change concepts). No introspective state is involved in this acquisition process. The concepts DANGEROUS and OFFENSIVE are acquired through perception (as are the relevant bodily-change concepts), while the concept EXPERIENCE is not acquired at all. Therefore, we do not need to posit a non-conceptual kind of introspective state to explain the acquisition of either FEAR or FURY. And what is true of these two may be true of all phenomenal concepts (including those associated with emotions or moods). Call this the *Hybrid Account* of phenomenal-concept acquisition.

Our reply is twofold. On the one hand, the Hybrid Account does not seem to fit naturally in Rosenthal's own view. On the other hand, the implausibility of some of its consequences gives a more general reason for finding the Hybrid Account unappealing. Let us consider these in turn.

We doubt Rosenthal himself would plump for the Hybrid Account, for at least two reasons. *First*, the account seems in tension with his view about qualitative properties of experience and the way humans pick out and discriminate them. The Hybrid Account fits naturally in a "transparency" conception of conscious experience (Moore 1903; Harman 1990, 1996; Tye 1995), on which we can be aware of properties of *objects* (e.g. the redness of a tomato), but not of properties of *experiences* (the reddish quality of an experience as of a red tomato). In such a framework, where introspective awareness never takes properties of experience as its object, but always "goes through" experience to what the experience represents, it is natural to explain the acquisition of phenomenal concepts along the lines of the Hybrid Account, i.e. by appeal to perception (awareness of objects and/or their properties) rather than introspection (intended as awareness of experiences and/or their properties). Rosenthal, however, rejects the transparency view. Against transparency theorists like Harman, he argues that properties of objects are not the only properties of which we can be aware: we can also, at least sometimes, be aware of the sensory or qualitative properties of experiences (Rosenthal 2000: 215-216). To be sure, one may reject the strong transparency picture (and thereby allow for the possibility of introspective awareness of properties of experience), while endorsing the Hybrid Account of phenomenal-concept acquisition. However, arguably, this is not the way Rosenthal himself might want to go, for this seems to be in tension with his own view about the way we pick out and discriminate experience's qualitative properties. He writes:

> When we classify sensory states and discriminate among their various tokens, we appeal to what it is like for us to be in those states. This is equally so with bodily and perceptual sensations; we rely on such things as what it is like to be in pain, and what it is like to see red or hear a trumpet. (Rosenthal 1991: 19)

Discriminating and classifying sensory states are functions carried out by phenomenal concepts. Arguably, *appealing to what it is like for us to be in a certain state* amounts to, or at least involves, *introspecting* the relevant state's phenomenal properties. Accordingly, it seems that, on Rosenthal's view, to acquire and deploy phenomenal concepts we need to be introspectively aware of phenomenal properties of our experiences– rather than perceptually aware of properties of the objects represented by those experiences.

Second, the Hybrid Account, when combined with Rosenthal's HOT theory of consciousness, has the implausible consequence that one cannot experience fear prior to a certain number of (fearless) episodes of awareness of danger. To see why this is a consequence, consider that on the Hybrid Account, the phenomenal concept FEAR can only be formed *after* having acquired the

perceptual concept DANGEROUS. To acquire the latter, though, one presumably needs first to (seem to) perceive danger a number of times. Therefore, perceptual awareness of danger is always prior to the formation of the phenomenal concept FEAR. Now, on Rosenthal's HOT theory, one can only *have* an experience of fear if one possesses the phenomenal concept FEAR (for the fear state is made conscious by a second-order thought that deploys that concept). By combining the Hybrid Account with Rosenthal's HOT theory, then, we have that it is impossible to experience fear without first perceiving danger (fearlessly) a number of times. But this is implausible. It seems infants could experience fear *before*, or at least as soon as, they perceptually represent danger. An infant may have a first experience of fear (perhaps immediately after birth!) even if she has never perceptually represented anything as dangerous. More generally, it is implausible that for any emotional experience, one can only have it after one has had prior perceptual experiences of such high-order properties as being dangerous, being offensive, and so on.

Beside its undesirable consequences for Rosenthal's own version of the HOT theory, there is a more general reason to reject the Hybrid Account. If the Hybrid Account were true, having a perceptual experience of danger (plus some suitable bodily sensations) would be sufficient for one who has never experienced fear to *simulate* or *imagine* a fear experience (in another or in oneself). For to simulate that, say, the zebra is afraid of the lion, one has to possess the concept of fear and imaginatively apply it to the zebra. According to the Hybrid Account, possession of the concept of fear does not require actually experiencing fear. It only requires innately possessing the concept of experience and having perceptual awareness of danger (and perhaps a few perceptual concepts associated with bodily changes). Thus a person who has never feared anything, but has undergone perceptual awareness of danger (and of certain bodily changes) would be able to "create" a new phenomenology in her mind—the emotional experience of fear—and to imaginatively apply it to someone else. This seems quite extraordinary.

For all these reasons, we think the Hybrid Account is inherently problematic and anyway not a good ally for Rosenthal's HOT theory (we doubt Rosenthal himself would be tempted by it). In addition, let us remind the reader that even if one accepted the Hybrid Account and concluded that primitive introspection need not play any role in phenomenal-concept acquisition, this would only deprive us of *one* theoretical reason for positing primitive introspection. It would not rule out primitive introspection's existence, which, as noted above, is strongly motivated by further considerations.

3.2. Defending (P2)

The HOT theorist may acknowledge the existence of (a non-classificatory introspective phenomenon akin to) primitive introspection, but deny that the TOT model is inconsistent with it. She may argue that, even though it does not accommodate primitive introspection as characterized in §2 (i.e., as a non-conceptual mental state), the TOT model *can* accommodate the existence of an introspective phenomenon which does not involve any classification or recognition of the introspected experience. What we have in mind are introspective thoughts that deploy *demonstrative* concepts, concepts such as THIS or THIS (KIND OF) EXPERIENCE, through which phenomenal characters can be picked out without quite being classified or even recognized. Appeal to such demonstrative concepts is made by conceptualists in other areas (McDowell 1994: 56-7) and could be used by the conceptualist about introspection to capture the kind of introspective phenomenon we are interested in, where a token experience is represented without being typed.

We find this approach more promising, but think it is ultimately not workable. Our argument against it is unfortunately not of the simplest kind—a trilemma nested inside a dilemma! At bottom, though, we will argue that the approach is ultimately forced into a completely implausible picture of the phenomenology of introspecting.

To present our argument against the demonstrative approach, we will need a conception of the content of demonstrative introspective thoughts. Now, there are different views one might have on this. On the one end of the spectrum, it might be a highly articulated sort of content, perhaps something like <I am having *this* kind of experience>, where 'this' denotes the relevant determinate experience type. On the other end of the spectrum, the content might be rather minimalistic – something like <this is occurring>, where 'this' refers to the relevant token experience. In-between a number of other options may be envisaged. Obviously, the more articulated the content, the farther one drifts from the kind of introspective phenomenon we isolated in §2. If the content of the introspective thought is <I am having this kind of experience>, for instance, then it would appear one would need to deploy the concept of self, the concept of a kind, and the concept of experience. For this reason, we will suppose for the sake of exposition that the demonstrative introspective thought appealed to in the TOT model has the content <this is occurring>. This supposition is unnecessary for our argument and plays only an expository role.

The problem for the TOT model is that it faces some uncomfortable choices with respect to the demonstrative concept THIS in <this is occurring>. In the TOT model we are considering, the first-order state M is made conscious by a second-

order thought with the content <this is occurring>, which is itself made conscious by a third-order thought, whose content is roughly <I think that this is occurring> (*Figure* 1).

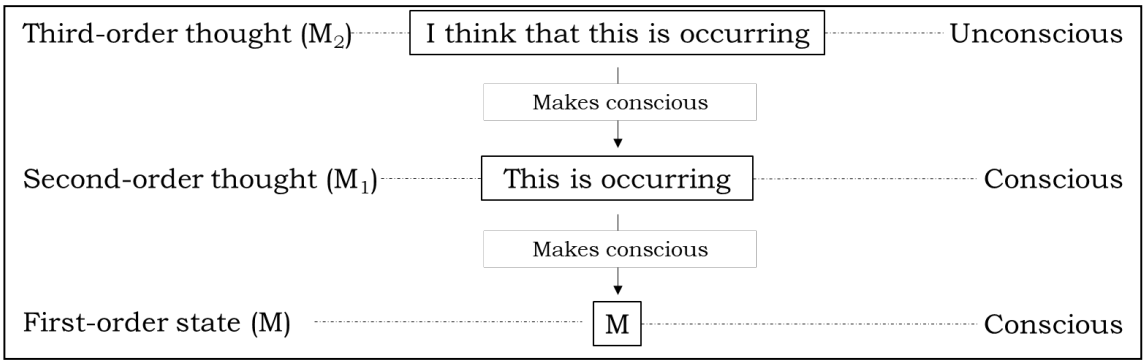

Figure 1

A first dilemma concerns whether THIS in <this is occurring> is a *blind* or a *substantial* demonstrative. A demonstrative is *blind* when it picks out its referent without carrying any information about it: it simply refers to 'this thing I am ostending (or attending to), whatever it may be.' A *perceptual* blind demonstrative, for instance, may be used by a subject who points blindly at a region of space before her, wondering what *that* is (Levine 2001: 82). Here the subject has no substantial conception of the referent of the demonstrative: no perceptual information about the referent is carried by the demonstrative—simply because the subject does not perceive what the demonstrative denotes. When the subject opens her eyes, perceptual information can fill in the content of the demonstrative, which thereby becomes *substantial*. (Perhaps there is not a *single* demonstrative concept that first is blind and then becomes substantial. Perhaps it is rather that a blind demonstrative is *replaced* by a substantial demonstrative. This will make no difference to our argument.) A demonstrative is *substantial*, then, when it carries *some* information about its referent.

Now, if there really are *introspective* blind demonstratives (not an uncontroversial issue: what would count as a "blind" mental pointing?), then the question arises whether THIS in <this is occurring> is blind or substantial. If it is blind, then the content of the demonstrative second-order thought amounts to something like <this (which I am mentally pointing at), whatever it may be, is occurring>. The demonstrative concept here denotes M (the first-order state, *cf. Figure 1*) without carrying any information about it. However, this is doubly problematic. First, this is manifestly not how introspection presents introspected experiences. When we introspect a pistachio-gelato experience, we are aware of some determinate qualities—that distinctive pistachio-y taste, for starters.

Introspection does not present a '*something, whatever it may be*.' Secondly, as we think of introspection (and as noted in §2), introspecting constitutes a kind of knowledge of the introspected. Arguably, this implies that, at the very least, through introspection one should get some *information* about the introspected mental state. But a thought featuring just a blind demonstrative and no substantial concept does *not* enable the subject to get any information about their current mental states. Therefore, it seems that, for the TOT model to do justice to the phenomenological reality and epistemic significance of introspection, some substantial concept must feature in the content of the second-order thought.[2]

What, then, are the prospects for a TOT model that marshals a *substantive* concept in higher-order thoughts? Our basic reason for rejecting this approach is that it has no stable way of deciding whether the information carried by the demonstrative featuring in the second-order thought is *phenomenal* or *non-phenomenal*. *Phenomenal* information is information about M's phenomenology, that is, information about what it is like for the subject to be in M (e.g. information about the reddish quality of one's visual experience as of a red tomato, or about the unpleasant feeling that comes with pain experiences). Quite obviously, for S to get information about M's phenomenology, M needs be phenomenally conscious: if M is unconscious, then there is *nothing* it is like for S to be in M and, a fortiori, S cannot get any information about the phenomenology of M. *Non-phenomenal* information is information about things other than M's phenomenology: information about properties M has independently of its being phenomenally conscious. The question is whether the substantial THIS featuring in the second-order thought carries phenomenal information or only non-phenomenal information about M.

Can the proponent of the TOT model construe the second-order THIS as carrying phenomenal information? This does not seem to fit the higher-order account of consciousness in the case of ordinary, non-introspective consciousness. In the non-introspective case, the second-order thought cannot carry information about *phenomenal* aspects of M, because M does not *have* phenomenal aspects independently of being higher-order thought of, in the first place. For the second-order thought to represent *phenomenal* properties of M, M

[2] It might be objected that primitive introspection, as characterized in §2, is threatened by a similar worry. How is a state that does not involve *any* concept supposed to provide one with *knowledge*? We only have the space here for gesturing toward the sketch of an answer. Roughly, the idea is that primitively introspecting constitutes a kind of knowledge that is *non-propositional* and *non-conceptual*—very akin to what Russell called "knowledge of things by *acquaintance*." The existence and epistemic significance of this kind of knowledge has been defended by some contemporary epistemologists (e.g. Hofmann 2014 and Duncan 2018; see also Giustina 2018 Ch.6).

would have to be phenomenally conscious. But the second-order thought is that *in virtue of which* M counts as conscious. The second-order thought cannot extract phenomenal information from M as long as M is not phenomenally conscious, and M is not phenomenally conscious until it is represented by the second-order thought, since the second-order thought is what *makes* M phenomenally conscious. It therefore seems that, at least in the non-introspective case, the second-order THIS must carry non-phenomenal information about M—if it is to carry information about M at all.

Can the TOT model construe THIS as carrying non-phenomenal information, then? Unfortunately, that does not seem to work in the case of *introspective* consciousness. Quite trivially, introspective states present *phenomenal* aspects of the introspected experiences. In the introspective case, then, the second-order thought—if that is what the introspective state is to be modeled as—represents some phenomenal properties of M. Therefore, plausibly, when the second-order thought <this is occurring> is conscious, the THIS carries *phenomenal* information about M.

Might the TOT model construe the relevant THIS as carrying phenomenal information in introspective cases and non-phenomenal information in non-introspective cases? This is coherent, of course, but has highly undesirable consequences. The content of the second-order thought <this is occurring> is partly *determined* by whether THIS carries phenomenal or non-phenomenal information. On the plausible assumption that concepts individuate sensitively to the kind of information they carry, two demonstrative concepts THIS_P and THIS_N that co-refer to M are *different* if the former carries *phenomenal* information about M but the latter carries *non-phenomenal* information about M. Accordingly, two thoughts featuring THIS_P and THIS_N respectively would have different contents and thereby be different thoughts (since thoughts are individuated by their content). If the second-order thought features THIS_N in the non-introspective case (when the second-order thought is unconscious) but features THIS_P in the introspective case (when it is conscious), then in becoming conscious the second-order thought undergoes an intrinsic change. That is, the occurrence of the third-order thought *changes* the second-order thought—it *does* something to it—where the relevant change is not a mere *Cambridge change*, but a change in intrinsic properties. However, as noted, on the HOT theory, a mental state does *not* undergo any intrinsic change when it becomes conscious. It is essential to the theory's reductive ambitions, we noted, that a mental state is intrinsically the same when it is unconscious and when it is conscious.

If all this is right, then the TOT model faces a destructive trilemma when it tries to account for primitive introspection by incorporating a substantial

demonstrative THIS into the second-order thought <this is occurring>. Either (1) the demonstrative always carries non-phenomenal information, or (2) it always carries phenomenal information, or (3) it carries non-phenomenal information when the second-order thought is unconscious (in the non-introspective case) and phenomenal information when it is conscious (in the introspective case). Each of these options, we have seen, has implausible or undesirable consequences. If (1) the demonstrative always carries *non-phenomenal* information, then the content of the second-order thought fits well with non-introspective cases but not in introspective cases. (1) has the implausible consequence that non-classificatory introspective states never represent the phenomenal properties of introspected states. If instead (2) the demonstrative always carries *phenomenal* information about M, then introspective cases are accommodated but non-introspective cases are not. (2) poses a circularity threat: if the second-order thought carries information about the phenomenology of M, then M needs *have* a phenomenology *prior to* the formation of the second-order thought, but on the HOT theory the second-order thought is exactly what *gives* M its phenomenology. As for option (3), it forces a radical revision of the HOT theory of consciousness. As noted in §1, on Rosenthal's view consciousness is a *relational* property, not an intrinsic property. Accordingly, when a higher-order thought makes a lower-order thought conscious, it does not *cause* a change in M's intrinsic properties. Consciousness is not an intrinsic property of M *caused* by the higher-order thought. Rather, it is a relational property *constituted* by the presence of the higher-order thought directed at the lower-order state. On a standard HOT theory, this is the case regardless of what "level" the higher-order and lower-order states are in. However, in endorsing option (3), the HOT theorist would reach the view that although a *second*-order thought does not change the lower-order state it makes conscious, a *third*-order thought does change the lower-order thought it makes conscious. In other words, although first-order states do not undergo an intrinsic change when they become conscious, second-order states do. The result would be a consistent but odd and unstable theory, according to which consciousness is a relational property of first-order states but an intrinsic property of second-order states.

This concludes our argument against the attempt to recover primitive introspection within a Rosenthal-style TOT model of introspection by adverting to demonstrative higher-order thought. As promised, the argument is a simple trilemma-nested-inside-a-dilemma. The structure of the argument is depicted in *Figure* 2.

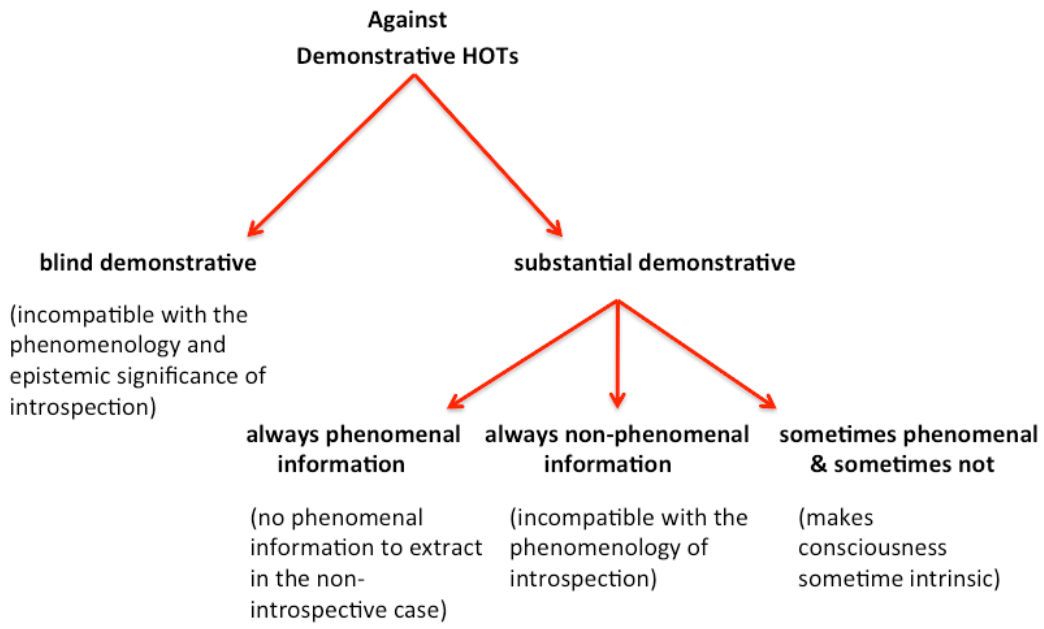

Figure 2

If the TOT model of introspection cannot *accommodate* primitive introspection, its only option is to *deny the existence* of primitive introspection. As we have argued, however, the psychological reality of primitive introspection is both manifest in everyday-life cases and theoretically indispensable in explaining the acquisition of phenomenal concepts. So, even if one accepted David Rosenthal's theory of introspection as an account of reflective introspection, the theory would still require supplementation in the form of an account of primitive introspection.[3]

References

- Duncan, M. 2018. "Knowledge of Things." *Synthese*. DOI: https://doi-org.ezproxy.rice.edu/10.1007/s11229-018-01904-0.
- Giustina, A. 2018. *Primitive Introspection*. PhD Dissertation, École Normale Supérieure.
- Giustina, A. 2019. "Introspection without Judgment." *Erkenntnis*. DOI: https://doi.org/10.1007/s10670-019-00111-8.
- Harman, G. 1990. "The Intrinsic Quality of Experience." *Philosophical Perspectives* 4: 31–52.

[3] We benefited from presenting this paper at the CUNY Cognitive Science Speaker Series. We thank the audience there, in particular Jared Blank, David Chalmers, Chad Kidd, Kate Pendoley, Adriana Renero, Kate Ritchie, and David Rosenthal.

- Harman, G. 1996. “Explaining Objective Color in Terms of Subjective Reactions.” *Philosophical Issues* 7: 1–17.
- Hofmann, F. 2014. “Non-Conceptual Knowledge.” *Philosophical Issues* 24 (1): 184–208.
- Levine, J. 2001. *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.
- McDowell, J. 1994. *Mind and World*. Cambridge, MA: Harvard University Press.
- Moore, G.E. 1903. “The Refutation of Idealism.” *Mind* 12 (48): 433–53.
- Rosenthal, D.M. 1986. “Two Concepts of Consciousness.” *Philosophical Studies* 49: 329-359.
- Rosenthal, D.M. 1990. “A Theory of Consciousness.” ZiF Technical Report 40, Bielfield, Germany. Reprinted in N.J. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness*. Cambridge MA: MIT Press, 1997.
- Rosenthal, D.M. 1991. “The Independence of Consciousness and Sensory Quality.” *Philosophical Issues* 1: 15–36.
- Rosenthal, D.M. 1993. “Thinking that One Thinks.” In M. Davies and G. W. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell.
- Rosenthal, D.M. 2000. “Introspection and Self-Interpretation.” *Philosophical Topics* 28 (2): 201–33.
- Rosenthal, D.M. 2005. *Consciousness and Mind*. Oxford: Oxford University Press.
- Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.