ABSTRACT

Title of the document: Consciousness and Mental Quotation: An intrinsic higher-order account

Vincent J. Picciuto, Doctor of Philosophy, 2014

Directed By: Professor Peter Carruthers, Philosophy

The guiding thought of this dissertation is that phenomenally conscious mental states consist in an appropriate pair of first-order and higher-order representations that are uniquely bound together by mental quotation. In slogan form: to be conscious is to be *mentally* quoted.

Others before me have entertained the idea of mental quotation, but they have done so with the aim of putting mental quotation to work as part of the "phenomenal concept strategy" (Papineau, 2000; Balog, Block, 2006; Balog 2012). Their purpose was importantly different from mine. According to those theorists, mental quotation is entirely introspective. On their views, a mental quotation is supposed to be a unique concept that we sometimes use to think about our own conscious states. Conscious states are assumed to be already conscious in virtue of some independent factor, or factors. Mental quotations are not supposed to be that in virtue of which conscious states are conscious. In contrast, this dissertation proposes that mentally quoting an appropriate first-order state is what makes a conscious state conscious in the first-place.

Treating consciousness as existing in a higher-order thought that mentally quotes first-order sensory contents has immediate explanatory dividends. It explains

several of the classic puzzles of consciousness as well as solving a set of puzzles to which existing higher-order theorists fail to respond. This includes what many see as an insurmountable problem for existing views: the problem of higher-order misrepresentation. If the higher-order component of a conscious state is quotation-like, the gap is filled between the state represented and the higher-order state that makes the state conscious. Rather than targeting a numerically distinct state from afar, as an extrinsic higher-order representation does, a mental quotation latches onto the very target state itself. The target state is enveloped and thereby becomes a component of the higher-order state, and it is the complex, the quotational state as a whole, that is the conscious state. What emerges from the guiding thought is a novel self-representational (or intrinsic higher-order) model of consciousness, described at the intentional level, which is immune to challenges facing existing views.

CONSCIOUSNESS AND MENTAL QUOTATION: AN INTRINSIC HIGHER-ORDER ACCOUNT

By

VINCENT J. PICCIUTO

A Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Paul Pietroski
Professor James Reggia
Professor Georges Rey
Assistant Professor Josh Weisberg

Table of Contents

Chapter 1: Introduction

This is the claim I elucidate and defend: each phenomenally conscious state is constituted by two components uniquely bound by mental quotation. The components are states of different representational orders: a first-order state representing some feature of the subject's environment, and a higher-order thought representing the lower-order state itself via a mental quotation. In other words, a mental state is conscious in virtue of the fact that it mentally quotes some or other sensory state, thereby incorporating that state into itself.

My claim, like other higher-order explanations of consciousness, is intended to be amenable to scientific explanation. How, then, might we test, confirm, or falsify the hypothesis, or any allegedly naturalistic hypothesis about phenomenal consciousness? Currently we don't know. Maybe we never will. It might turn out to be true that a complete grasp of consciousness is beyond the limits of our cognitive capacity. I am not, however, ready to cast my lot with mysterians just yet. At this point in the inquiry, much of the theory-building to be found in the consciousness studies literature is speculative, for much of what we can do is to hypothesize, argue *how possibly and how plausibly*, and *begin* to look for supporting data. These activities are worthwhile, given the current state of the field, and state of the art in cognitive science. A plausible intentional model tells us where to begin when we develop a computational model. It can provide the logical structure such a state might have, is likely to have, and cannot have. That is what I do in this dissertation. I spell out the notion of mental quotation at the intentional level and then explain the possible implications it has for a theory of phenomenal consciousness.

The primary result is the demystification of some of the initially puzzling data with which we are confronted when we reflect on our phenomenally conscious experiences. Additionally, we can avoid some of the main problems facing traditional higher-order theories and existing intrinsic ones. We can do this without elaborate machinery and arcane commitments. But we must warm to our subject. That is the aim of this introductory chapter. In this introductory chapter, I will draw preliminary distinctions and lay out the assumptions that I will be working with throughout this dissertation.

In section 1, I briefly characterize phenomenal consciousness, *i.e.,* the explanandum of the dissertation. In section 2, I highlight the specific puzzles that will be the focus of subsequent chapters. In section 3, I review the key features of the representational theory of consciousness and some of its main challenges, none of which count decisively against the view. In section 4, I discuss the phenomenal concept strategy, from where the idea of mental quotation is drawn. I do this to distinguish my mental quotation account of consciousness from the role of mental quotation in the phenomenal concept strategy. In section 5, I discuss the distinction between what are known as first-order representational theories (FOR theories) and higher-order representational theories (HOR theories). In that section I reemphasize one main argument against FOR theory.

As this is an introductory chapter, the central claim of the chapter is modest: the representational theory is a plausible strategy. However, while first-order theory, in particular, lays the groundwork for that strategy, it has critical shortcomings which motivate a higher-order view (Carruthers, 2000, 2005; Kriegel, 2006; 2009). My brief assessment of the FOR theory should not be taken as contributing a wholly original argument against first-order theory. I intend it to be largely a reemphasis of what others

have already argued, in the way of providing some theoretical background for the higher-order approach that will concern the rest of the dissertation, and for the novel approach to be introduced in Chapter 4. In Chapter 2, I review three main problems for extrinsic HOR theory, arguing that one in particular motivates an intrinsic higher-order approach. Then, in Chapter 3, I evaluate existing intrinsic HOR theories, arguing that none adequately addresses the main problem they set out to address. This is the problem of higher-order misrepresentation. This sets the stage for the novel intrinsic model that is develop in Chapter 4. In Chapter 5 I discuss possible objections and future directions of the view.

1. What Are We Even Talking About?

Conscious experiences have a mental appearance. They *seem* certain ways *to us*. They are, as the expression goes, "like something" for us (Nagel, 1974). When you see the blue sky, there is something it's like for you. There is some "feel" the experience has that is different from your conscious experience of a pink sky or a toothache. But such experiences are never quite as simple as the examples philosophers tend to use. When you walk out to Route 1 and see the moving cars, hear the moving cars, see a green traffic light, barely distinguish random conversations, bump into someone passing by, noticing the blue sky, among other things, at the same time or in rough simultaneity, there is a way that entire complex experience "feel."

Distinguishing phenomenal consciousness from access consciousness, Block (1995) called consciousness a "mongrel concept." In fact, phenomenal consciousness and access consciousness are only two among the list of distinctions that often begin discussions of "consciousness," (*e.g*. phenomenal consciousness, access consciousness, subjective

consciousness, state consciousness, creature consciousness, introspective consciousness, self-consciousness, qualitative consciousness, narrative consciousness). Some of these proposed kinds are synonymous, some are merely terminological distinctions, whereas others mark principled distinctions. Some are firmly established, others more contentious. Presently, I will not discuss the many things one might mean by 'consciousness'.[1] The issues on which this dissertation is focused all arise from phenomenal consciousness, and all of the theories to be considered are attempts to account for it in particular. The interesting things about states like those described above are the ways the qualitative components seem to us *when we're consciously aware of them*. My understanding of phenomenally conscious awareness, and the one that I shall employ throughout, is this: phenomenally conscious awareness is just the kind of awareness exhibited by states with "phenomenal character."[2] Phenomenal character requires qualitative character (at the

[1] NB: throughout I use both single quotes and double quotes in different instances. I will be employing a method that Davidson (1979) and others had once tried. Single quotes literally mean *the expression,* whereas double quotes indicate something like *this is a new term that hasn't yet appeared in the text and is being introduced*, or *so-and-so said* the enclosed expression, or *the enclosed expression is being used in a funny way*. For example, the expression 'consciousness' means *the expression consciousness,* whereas the expression "conscious" can indicate something like *'conscious' is being introduced into the text for the first time,* or that so-and-so uttered the enclosed expression, or that '*conscious' is being used in a funny way,* as in "My cat seems quite dumb. I suppose maybe it's "conscious," but I doubt it's really conscious." To simplify things I have ignored the possible distinction between quoting "signs" as opposed to "expressions," which is discussed in Chapter 4 and in Cappelen and Lepore (2006).

[2] Different theorists use different terminology to pick out this feature. Some other common expressions are "raw," "subjective," or "phenomenal" *feel*. "Subjective character," "phenomenal character," and "qualitative character" are also sometimes used synonymously. Following others before me, below I will distinguish between qualitative and phenomenal character. My terminological preference throughout is to say that phenomenally conscious experiences have phenomenal character, or a "what-it's-like" (sometimes "what-it's-likeness) feature.

first-order) and subjective character (at the higher-order). Together, the two main components comprise the state's phenomenal character.[3] Such states that have it are phenomenally conscious, and generate what some would refer to as "the hard problem" (Chalmers, 1996).[4] Nevertheless, I do want to highlight a triad of the most significant standard distinctions that turn out to have theoretical consequences, and that will be directly relevant to various parts of the dissertation.

First is the distinction between mental *state consciousness* (a property attributed to a mental state) and a different, more common, sense of 'consciousness'.[5] This latter common conception of consciousness is *creature consciousness* (a property attributed to an organism), which, put roughly, amounts to an organism being awake, not being comatose, or perhaps, not sleeping dreamlessly, as for example, when we say that someone "lost," "regained," or "is coming in and out of," consciousness. The allegedly mysterious kind of consciousness (the property of being phenomenally conscious) that is thought to pose explanatory problems for physicalism, and that is the focus of this dissertation, is importantly different from merely being awake. With few exceptions (*e.g.,* Lau and Brown forthcoming), when theorists speak of *the* problem of consciousness, they

---

[3] This characterization of the three main components of phenomenally conscious experience derives from Kriegel (2009).

[4] Is the hard problem really the *hardest* problem? The claim that it is the hardest problem is looking increasingly contentious. Hereafter I will refrain from using 'hard problem' altogether. For it might be that specifying how brain states can manage to represent in the first place, independently of questions about consciousness, is the really hard problem.

[5] NB: in this sentence italics are functioning as double quotes; they're introducing a technical term that hasn't been introduced yet.

are referring to the problem of phenomenal consciousness as a species of mental *state consciousness*, in particular.

The second distinction is between what philosophers have dubbed *transitive* consciousness and *intransitive* consciousness. The notion of creature consciousness introduced above was implicitly intransitive. That is, intransitive creature consciousness is something akin to being awake, but an intransitively conscious creature (one that's awake) can also be aware *of* various things. This notion of being aware *of* something or other is the notion of *transitive* (creature) consciousness.

The last distinction I want to highlight is the distinction between phenomenal consciousness and *access* consciousness (McGinn, 1982; Davies and Humphreys, 1993; Block, 1995).[6] As stated above, 'phenomenal consciousness' refers to the what-it's-like feature of conscious states. 'Access consciousness', on the other hand, refers to the availability of certain contents, or mental states, to guide rationally action and verbal report. For example, it is not difficult to envision an information processing system that is capable of language production that is driven by states that are access conscious, and which may be entirely specified in functional terms. However, most would find it controversial to attribute phenomenal consciousness to such a system, merely in virtue of the availability of such states to guide action rationally. There seems to be something important that is lacking. Whether or not there really is something important lacking remains to be seen as our understanding of phenomenal consciousness grows. While not

[6] While the first two distinctions are firmly established, not everyone acknowledges the distinction between access and phenomenal consciousness. For an instructive critique see Church (1998).

all agree that these distinctions pick out genuine phenomena, throughout this dissertation we will see the roles they play in the different theories discussed.

2. Some Core Issues

The "hard problem" of consciousness is typically characterized by several puzzling features.[7] I cannot hope to address, let alone discuss, all of them here. I will focus on the following (D1-D5). I construe these as data with which we are confronted when reflecting on conscious experience, and about which it would be useful for a theory of phenomenal consciousness to have something to say.[8]

D1. What-it's-likeness, Subjective Feel, And Other Related Expressions

As mentioned above, different theorists use different expressions to refer to the allegedly mysterious kind of consciousness (what-it's-likeness, subjective feel, &c.). Whatever we call it, this elusive phenomenon is the essential feature of phenomenally conscious experiences. As such, any theory of consciousness should improve our understanding of it. It is, in fact, one of two features that I will claim are adequacy conditions for a theory of consciousness. Any theory of consciousness stands or falls on

[7] See, for example, (Tye, 1995; Chalmers, 1996; Levine, 2000; Carruthers, 2000, 2005).

[8] Some might think these data (in addition to others that I do not address) do in fact constitute necessary and sufficient conditions. However, I am not going to constrain the explanandum prior to theorizing about it. We currently don't know what phenomenal consciousness really is, and it *might* turn out, through the course of theorizing, that some features that seem obviously part of it aren't and that some that seem unrelated are.

whether it can explain (or explain away) what phenomenal feel consists in, even if the ultimate conclusion is that it is merely a mental appearance.

D2. The Conscious/Non-Conscious Distinction

Some mental states exhibit what-it's-likeness, while others do not. In what, exactly, does the difference between these two kinds of state consist?

There is now a large body of data that has amassed in support of the claim that we undergo non-conscious representational states (including perceptual states) and that such states play a significant role in the production of much of our action (construed broadly to include thought).

One datum that has received much discussion is the dorsal-stream visual system investigated by Milner and Goodale (1995) and others. This system seems to be a genuine visual system, even though the subject *can never* be aware of percepts along that stream. This system shares the same receptor-organ (the retina) as the ventral-stream system that is responsible for conscious vision, and it shares some of the same mechanisms immediately thereafter. But the two streams diverge soon after V1, and have very different processing properties, and quite different functions. The conscious system informs judgment and planning and deals in allocentric spatial representations, whereas the dorsal system is charged with the fine-grained on-line control of action and uses representations of space that are body and limb-centered. Importantly, there seems no difference between the two in the extent to which they qualify as *sensory* systems: most notably they each generate (sensory) representations (partly non-conceptual representations with mind-to-world direction of fit), it is just that the former sensory

representations are inaccessible, while the latter are sensory representations that are accessible (and/or accessed).

One way to characterize the difference between the two kinds of representation is by appeal to the qualitative/phenomenal distinction. Not all theorists acknowledge the distinction. However, there is good reason to do so. As just stated, the first-order perceptual representations along the dorsal stream seem just as sensory as those along the ventral stream. A sensory state such as a visual perception of color represents (at least) surface reflectance properties. But surface reflectance properties are paradigmatically "qualitative." Since those states actually figure in guiding our action, it looks like those qualitative properties are in fact represented in non-conscious perception. One anecdotal example is absent-minded driving (Armstrong, 1968). In such cases the driver is thought to perceive non-consciously, say, a red light, which may cause her to step on her brake pedal. There are various data that have emerged from controlled studies as well.

One well-known study that has generated much discussion involves blindsight subjects (Weiskrantz, 1986). Blindsight subjects have had injuries to their primary visual cortices, and they cannot consciously see objects that are presented in the visual field that corresponds to the injured area. However, the interesting thing about blindsight subjects is that they can non-consciously perceive some of the qualities of objects in their blind spots. That they can do so is indicated by the frequency and regularity with which they guess correctly when questioned about the relevant objects and locations, in some cases involving precise positioning (Marcel, 1998; Ramachandran and Blakeslee, 1998).

Blindsight studies are somewhat contentious and have been used to establish competing conclusions. For example, both FOR theorists and HOR theorists claim to

have plausible interpretations of blindsight cases. Prinz (2012) takes an altogether different route, arguing that blindsight studies suggest that subjects cannot recognize *objects* in their visual fields, even non-consciously. They do not represent objects as such. Rather, Prinz argues that their capacity to successfully guess locations "probably involves subcortical structures and, perhaps, a select subset of spatially sensitive cortical visual areas" (2012, 81). So while blindsight might establish that *some* features are non-consciously perceived (spatial locations). They might not establish that all features can be. Given the existing disagreement, it is helpful to look for additional examples.

Another oft-cited example of nonconscious perception is found in the literature on semantic priming. Under some conditions we seem to perceive words in the absence of consciousness, non-consciously recognizing, say, semantic relations between some but not others (Meyer and Schvaneveldt, 1971; Dehaene *et al,* 1998). For example, in what's known as a "lexical decision task" subjects are presented with a mixture of words and non-words as targets and their task is to indicate whether the target is a word or not. Preceding these targets are non-conscious primes, which can either be semantically related (doctor - nurse) or unrelated (butter - nurse) to the target words. In this kind of task the priming effect is expressed as a faster and/or more accurate response to semantically related prime-target pairs compared to unrelated pairs.

Perhaps the best example of robust non-conscious perception comes from research on unilateral neglect. Unilateral neglect is basically a disorder resulting from injuries to the right inferior parietal cortex (Driver and Mattingly, 1998; Driver and Vuilleumier, 2001). People with unilateral neglect seem to lack conscious experience in their left visual fields and/or of the left sides of objects. One telling example comes from

Marshall and Halligan (1988). They presented a neglect subject with pictures of two vertically aligned houses. In the pictures the houses are identical with one exception: one of the houses had flames shooting out of its left side. The subject reports that both houses are visually identical, but when asked which one she would rather live in, she consistently (nine out of eleven times) chose the one without the flames. Doricchi and Galati (2000) replicated the results. Their subject chose the flameless house seventeen out of nineteen times, while claiming that the two images were identical.

One last case is worth mentioning. It combines neglect and semantic priming. Berti and Rizzolatti (1992) illustrated that information in the neglected field can actually influence semantic priming. Subjects were asked to view pictures with pairs of objects (either fruits or animals) on their blind sides. They observed that subjects were able to categorize a visible object on the right side, if there was an "invisible" item from the same category on the left side.

What should we say about the kinds of contents that are constituents of these experiences? They clearly involve representations of perceptual properties: color (the redness of the traffic light or fire), shape (the curve of an 'S'), or sound (the sound of an 'S'). These are paradigmatic qualitative properties functioning in ways that directly influence action. There is every reason to consider them "experiences." However, since the subject is not aware of them, they are *non-conscious* experiences.[9] One task for a theory of phenomenal consciousness is to account for the distinction between these two kinds of state.

[9] Cf. Block: "Phenomenal consciousness just is experience" (1997, 277)

An account of D1 (subjective feel) would presumably tell us how to distinguish between non-conscious and conscious states. However, it should be noted that an account of D2 would *not* necessarily provide an account of D1. For example, there might well be differences between the causal path leading up to a nonconscious state and the causal path leading up to a conscious one, such that the two could be meaningfully distinguished. That would provide some account of the distinction between the two kinds of state, but it would still fail to explain what consciousness itself consists in. It would still fail to explain why the conscious path is the conscious path. Any theory of consciousness must say something about D2 *while being able to explain* D1.

D3. Intimacy

It is often claimed that we are intimately related to our own conscious experiences, but it is not always clear what intimacy claims are intended to assert. In fact, there is a family of notions traveling under 'intimacy', the members of which should be kept largely distinct. Typically, though, the apparent intimacy of conscious states is thought to be a problem for representational theories. I will briefly sort through some of the principal notions that one finds in the literature and then review two very general ways that one can handle all of them.

Sometimes the intimacy of conscious states is thought of as "immediacy," or as being the result of "unmediated" processes. But 'immediacy' itself is ambiguous between two distinct senses. There is a temporal sense of 'immediacy' and there is an epistemic sense as well. For example, Kriegel (2006) argues that there can be no temporal distance between the state that realizes an experience and the state that realizes our awareness of

that experience. The causal path from one to the other must somehow be directly simultaneous. The path must be unmediated by any intervening temporally extended process. On the other hand, there is also the sense of "immediacy" according to which conscious states are generated in a way that is "direct" as opposed to as the result of inference. Here 'direct' means something like noninferential.[10] For example, it is commonly held that in situations wherein one discovers that one is undergoing a state, say, on the basis of being convinced by one's therapist, that such a state would not thereby become phenomenally conscious.

To make matters more confusing, there are two distinct but relevant senses in which conscious states are typically characterized as being noninferential. On the one hand, there is the sense in which conscious states are not *generated* by an inferential process and, on the other hand, there is the sense in which what it is like to undergo a conscious state cannot be *known* merely in virtue of an inferential process without having undergone the relevant conscious state (this latter sense characterizes the situation of Jackson's now-famous Mary while she is in her black and white room and deserves to be treated as a separate datum, to be discussed below as D4).[11] Not only can't you get to be in a conscious experience via inferential processes, according to the latter sense of 'noninferential', you also cannot know what a specific conscious experience would be like to be in via inferential processes.

---

[10] See Rosenthal (2005) for an example of the noninferential notion of being (or at least seeming) unmediated. Though, one should note that Rosenthal seems to think that genuine inferences are always conscious.

[11] See Jackson (1986).

There is yet a further dimension of immediacy that pertains specifically to higher-order theories, and which depends on issues that have not been discussed yet. This will be dealt with in Chapter 2. No matter which sense one has in mind, though, there are two general strategies one might adopt to deal with the presumed intimacy of experience. One might claim that intimacy in the above mentioned senses is genuine and attempt to account for it (Carruthers, 2000; Kriegel, 2005, 2009). Or, one might think that it is only apparent, and attempt to explain it away (Rosenthal, 1990, 1993).

D4. Phenomenal Knowledge

The latter sense on 'noninferential' that was just discussed (the sense in which one cannot know what it's like to be in a phenomenally conscious state by inference alone) has received much attention and deserves to be treated as its own separate datum. This was the special kind of epistemic relationship that we seem to have with our own conscious states, and which can be illustrated by the now famous case of Mary, from Jackson's (1986) thought experiment. While the thought experiment was originally proffered as a strike against physicalism, it should be treat as another relevant datum with which we are confronted when we reflect on conscious experience. Surely most agree that when Mary leaves her black and white room and exclaims "*That's* what it's like to see red!" there is in fact *something* interesting that is different about Mary. This "something," I take it, is phenomenal and it requires explanation.[12] However, we can

[12] Originally the phenomenal was thought to be troubling, construed to imply an antiphysical, perhaps non-relational, intrinsic, and therefore unanalyzable component. Thus, we find Lewis (1983, 1988) rejecting so-called "phenomenal information." Here, what I mean by 'phenomenal' is just the what-it's-like, or mental appearance, component

admit that Mary is different after consciously seeing red while also rejecting the antiphysicalist conclusion. To do this, it has become common for physicalists to appeal to the "phenomenal concept strategy." The phenomenal concept strategy is intended to account ofr Mary-type cases, zombies, and inverts without denying physicalism. But, if one thinks that unification is worth striving for in scientific explanation, then if a theory of consciousness *itself* can help us understand this interesting feature, that would strengthen the theory.[13]

3. Physicalism and The Representational Theory of Mind

The preceding characterizes some of the initially puzzling features of consciousness. How might one respond to these initially puzzling features? There are, broadly speaking, three ways. First, one might argue that consciousness cannot be explained naturalistically, either because consciousness is constituted by at least some non-physical features, or because consciousness is beyond our cognitive reach. According to the former dualist approach, consciousness essentially involves at least some non-physical components, and thus, it will always evade a purely physicalist explanation (Chalmers, 1996; Gertler, 2006; Strawson, 1994). According to the latter "mysterian" line of thought, consciousness (whether it is physical or not) is ultimately mysterious relative to human cognitive capacities; we don't understand it now and we never will, even in the limit of scientific enquiry (McGinn, 1989).

---

discussed above. Importantly, though, this is *not* to concede anything to the antiphysicalist, nor to inflate one's ontology. For, as I hope to make clear, and as I think others have already made clear, the phenomenal can be construed in physicalist terms.

[13] The phenomenal concept strategy will be discussed in section 4.

Second, one might argue from within a physicalist framework that there is no, or might not be any, real or theoretically interesting property to be explained in the first place. This is more or less the eliminativist materialist approach. There are various ways to develop an eliminativist approach to consciousness (see, *e.g.,* Dennett, 1978, 1988; Hardcastle, 1999; and Rey, 1983, 1988). At its core, though, eliminativism amounts to explaining away phenomenal consciousness altogether.

One needs to be clear, however, on just what one is eliminating. It's implausible, if not impossible, to eliminate mental appearance itself. The elimination of "qualia," on the other hand, in the most inflated sense is at least initially plausible. But if that's all that is meant by 'eliminativism', it's possible that eliminativism is being conflated with reductionism of some kind. Some have mistaken reductionism for eliminativism.[14] But one can deny that there are mysterious, unanalyzable, non-physical phenomenal properties, while at the same time acknowledging that there is a real property in which our mental appearance consists that might be explained naturalistically. In other words, one can deny that there are qualia, in the most inflated sense, without denying phenomenal consciousness altogether.

Moreover, one might try to give a reductive *explanation* without ontologically reducing phenomenal consciousness. That is, one might try to explain phenomenal consciousness in non-phenomenal terms while not reducing the phenomena ontologically to neurons firing, or quantum particles. To do so would constitute a third possible response: to provide a physicalist, non-eliminativist account of phenomenal consciousness. One promising reductive approach is the representational theory of

[14] See Lycan, W. and Pappas, G. (1972) for clarification of this confusion.

consciousness and one of my main goals is to increase its plausibility. However, as I will make clear throughout, existing versions face substantial problems, leaving aspects consciousness unexplained.

Physicalists claim that mental states just are brain states, or that they are realized by brain states. The brain is, of course, a physical object, extended in space-time. On a physicalist account of consciousness, then, one might intuitively assume phenomenal consciousness should be a property of brain states. Now, if one goal of physicalism in the philosophy of mind is to provide a physicalist account of phenomenal consciousness, which is a property of mental states, and mental states (and their associated properties) are brain states/properties, then one might reasonably assume that neurobiology will be the appropriate level to explain accurately phenomenal consciousness. For example, one might supplement the phenomenal concept strategy with a type-identity version of physicalism, offering an account of phenomenal consciousness in terms of its putative neural correlates. That seems like the obvious explanatory path for physicalists, and it's the path that identity theorists of the 1950's and 60's followed (*e.g*., Smart, 1959), and the path that some argue physicalists must still follow (*e.g.,* Kim, 1992; 1998).

As stated above, the approaches that I will focus on in this dissertation are physicalist, but they are *representational* theories. The main difference between a neurobiological theory and a representationalist theory is that representational theories operate at the cognitive/functional level of explanation, not at the neurobiological level. It is not that representationalists think that mental states are not realized by neurobiological states in typical humans, but rather that the appropriate level of explanation for consciousness isn't neurobiology. The main reason for this is the following. First,

neurobiological theories will most likely be "type-identity" theories, claiming that each specific kind of mental state is identical to a specific kind of neurobiological state. It is, however, plausible to think that *some kind* of "multiple realizability" is possible. And if so, that makes strict type-identity problematic. This is hardly a refutation of type-identity theory, but it's not meant to be. All it is meant to do is hint at one general line of reasoning against the view. Representationalists, will most likely be "token-identity" theorists. Token identity theorists claim that a given mental state is identical to some neurobiological state or other, but that mental states can be multiply realized, if not only by distinct kinds of neurobiological state, but also, perhaps, by distinct kinds of physical states, *i.e.,* states other than human brain states (silicon states, aluminum cans and strings, light bulbs, or whatever).

It might well be that certain materials set functional constraints, though. If so, the idea of conscious brains comprised of tin cans and strings can probably be ruled out. That is, multiple realizability for mental states may not be quite as easy to come by as the standard functionalist arguments suggest, and so the relationship between form and function, in this case the relationship between neurobiological structure and function, could feasibly be tightened.[15] No matter what the neural correlate of consciousness is—moreover, no matter what the neural correlate of any kind of contentful mental state is, according to the representationalist, what is most interesting about such states for the purposes of psychological explanation is the content of such states and the causal/functional role they play in one's cognitive system. Neural correlates are surely

[15] Prinz (2012) develops this line of "neuro-functionalist" argument, challenging the traditional contrast between identity theory and functionalism.

relevant to psychological explanation. If token-identity is true, however, they do not on their own provide an exhaustive explanation for all psychological phenomena, and in particular, for phenomenal consciousness, especially at this stage in the development of neuroscience.[16]

Representationalists are motivated, in part, by the above worries with type-identity. For all naturalists, the main goal is to give an explanation of phenomenal consciousness in non-phenomenal terms. But this does not require that we leap from high-level phenomenal terms directly to low-level neurobiological ones. According to representationalists, the appropriate lower-level at which to explain phenomenal consciousness in non-phenomenal terms is the level of representational content and the functions that such contents have within the causal structure of one's cognitive system.

Much of the plausibility of the representational theory of consciousness (RTC) is no doubt due to the plausibility of the more general representational theory of mind (RTM). The idea that the mind is representational has a rather long pedigree, and it is consistent with the claim that mental states are not physical (*e.g.,* Aristotle, Descartes, Locke, Berkeley, Kant, Hume, Brentano and many others all seem to have thought that mental states represent the world to us, or exhibit intentionality, or are about things, but many of them also thought that such states were not physical states).[17] Only during the

[16] For example, we will want to know whether certain cognitive functions that one posits, in which contents and/or their vehicles are supposed to play some causal role, can be implemented by the brain and whether they actually are. That's one task of cognitive neuroscience.

[17] Strictly speaking, what I am calling the "representational theory of mind" is what some would once have called the "computational-representational theory of mind." However, it has become common parlance in the consciousness studies literature simply to assume

past century has the idea that the mind is representational been proposed as a physicalist theory of the mind, and even more recently as a physicalist theory of consciousness in particular.

'Intentional' is initially idiosyncratic, but the idea that mental states are about things is straightforward. For example, to believe occurrently *that Federer lost last night* is to undergo a mental state (a belief that Federer lost last night) that is about the result of a particular tennis match involving Federer. My desire for cold milk in a warm glass, straight from the dishwasher is about cold milk and a warm glass. These states are "representational." They represent things in the world to me, *e.g.,* Federer, cold milk, and a warm glass. Historically, though, such states were not necessarily thought to be physical. On the other hand, according to current conceptions of RTM, these states are non-mysterious physical states of the brain that bear content.[18] If it succeeds, RTM would provide one possible path of naturalization for mental content. This forms what is perhaps the main motivation for the representational theory of *consciousness* (RTC).

Proponents of RTC argue that the theory promises to naturalize phenomenal consciousness by reducing phenomenal consciousness to representational content (and

---

that RTM, as a reductive thesis, involves a computational component, even though non-reductive RTM is consistent with anti-physicalism. So, for ease of use, I will omit the prefix 'computational' and continue to employ 'RTM' and 'RTC', instead of 'CRTM' and 'CRTC', which some might prefer (*e.g.,* Rey 1998).

[18] Among the competing naturalistic theories of mental content are: informational views (Stampe, 1977; Dretske, 1981; Fodor, 1987, 1990, 1991), conceptual role views (Field, 1977; Loar, 1981; Harman, 1982; Block, 1986), and teleological variations of each (Milikan, 1984; Papineau, 1987; Dretske, 1988; Neander, 1991; and Godfrey-Smith, 1994, 1996).

causal/functional role), which itself can be explained naturalistically. [19] Even though there are many competing theories that aim to naturalize mental content, each of which have their own pitfalls, the possible paths to naturalization are at least in view for RTM. While many will object to the very idea of naturalizing phenomenal consciousness right from the start, if we can reduce phenomenal character to representational content, then we are one step closer to naturalizing phenomenal character.

4. The Phenomenal Concept Strategy

Physicalists of any kind, whether representationalist or not, have a plausible way of meeting a range of anti-physicalist challenges (for example, well-known arguments such as the knowledge argument, Jackson, 1986; explanatory gap arguments, Levine, 1983, 2001; and zombie/invert arguments, Block, 1978, 1990; Chalmers, 1996). These arguments present physicalism with epistemic puzzles that have purported metaphysical consequences, *viz.,* they are supposed to imply that physicalism as a metaphysical thesis about the nature of phenomenal consciousness is false. One response to these challenges is what is known as the "phenomenal concept strategy." [20]

---

[19] Some might think the above parenthetical is contentious. For example Kriegel (2002) argues that the PANIC theory is not a genuine representationalist theory but rather "disguised functionalism." I will elaborate on the merits of this claim below. As a preview, it is not clear that anybody at all attempts to offer a reductive representationalist theory without any appeal whatsoever to the specific functions of representational content.

[20] Stolar coined the phrase 'phenomenal concept strategy' (Stoljar, 2005). It has been endorsed by others who reject the quotational account. See for example, Loar's (1990), Carruthers' (2000), and Tye's (2000) recognitional accounts, and Perry's (2001) and O'Dea's (2002) indexical accounts.

The main thrust of the phenomenal concept strategy is to explain away this range of challenges to physicalism by attributing the apparent mysteriousness of phenomenally conscious states to the way we sometimes think about such states. Proponents of the strategy claim that in many cases our thoughts about our own phenomenally conscious states are constituted by phenomenal concepts, which are unlike any other kind of concept. Such concepts, according to those who embrace the phenomenal concept strategy, can be given a physicalist explanation. If successful, the strategy shows that it is the purported distinctness of phenomenal concepts (not consciousness itself) somehow gives rise to the puzzling features put to work in the anti-physicalist arguments. Because such concepts can be explained naturalistically, the anti-physicalist arguments fail.[21]

For example, given the distinctness of phenomenal concepts (sometimes phrased in terms of "conceptual isolation") proponents of the strategy argue that it is no surprise that, say, Mary—expert in the neuroscience of color vision in Jackson's knowledge argument — learns something new upon stepping out of her black and white room and consciously perceiving red. Before leaving her room Mary lacked the relevant phenomenal concept. Consequently she could not derive knowing what it's like to see red from her complete physical description of red. When she left the room and saw red, she gained the relevant phenomenal concept, enabling her to know what it's like to see red. Essentially, proponents of the phenomenal concept strategy argue that consciousness itself is not mysterious, but the way we *think about* our own conscious states makes it

[21] It is well-known that Jackson has given up his original presentation of the knowledge argument as refuting physicalism. Also, Levine's explanatory gap argument was never supposed to *establish* the falsity of physicalism. Nevertheless, these arguments are often presented as either defeating *or* presenting a puzzle for physicalist views of phenomenal consciousness.

seem like consciousness is mysterious. Thus, Mary's new knowledge, explanatory gaps, and the conceivability of zombies and inverts are not surprising nor ultimately threatening to physicalism. The main task for proponents of the strategy is to explain the distinctness of phenomenal concepts in a way that is consistent with physicalism. There are different ways that philosophers have tried to do this. For example, Loar (1990) and Carruthers (2000, 2005), and Tye (2000) argue for "recognitional" accounts. Perry (2001) and O'Dea (2002) argue for "indexical" accounts. Papineau (2000) and Balog (2008) argue for a "quotational" account. The quotational account is the one on which I will focus in this dissertation. As will become clear in Chapter 4, however, my understanding of mental quotation and its role within cognition is quite different from Papineau's and from Balog's.

Whichever account of phenomenal concepts one chooses, the phenomenal concept strategy itself makes no mention of any specific physicalist account of phenomenal consciousness. In fact, one of the pluses of the strategy is thought to be how it distances an explanation of consciousness from an explanation of how we think about consciousness. It relegates the initially puzzling data to the latter project. Thus, the general approach that emerged over the past two decades or so presents a two-pronged solution to the problem of consciousness: (i) the phenomenal concept strategy, plus (ii) some or other particular physicalist account of consciousness itself.

On the one hand, the phenomenal concept strategy would provide an account of the special way we sometimes think about our own conscious states, which in turn is supposed to provide a response to the range of critical anti-physicalist arguments. On the other hand, a specific physicalist account of phenomenal consciousness would tell us

what phenomenal consciousness itself is. According to the received view, the two projects are largely independent, and the phenomenal concept strategy is supposed to be consistent with any of the main variants of physicalism (except for ones that outright reject the strategy, *e.g.,* Tye 2009, or perhaps, for views that fail to account for, or are inconsistent with, the kind of concept required by the phenomenal concept strategy).[22]

The phenomenal concept strategy itself figures in this dissertation mostly because the mental quotation account of consciousness that I develop in Chapter 4 has its origins in the quotational account of phenomenal concepts. Whether or not the mental quotations that I put to work in Chapter 4 are the very same phenomenal concepts at work in the phenomenal concept strategy is an issue I will discuss Chapter 5.

5. From First-Order to Higher-Order Representational Theories

There is one major division among representational theories of consciousness. This is the division between first-order representational theories (FOR theories) and higher-order representational theories (HOR theories). According to FOR theorists, a mental state M is conscious if and only if it represents appropriately. There are different ways to cash out representing "appropriately." One thing that all FOR theorists have in common, though, is a commitment to the claim that higher-order representational, or metarepresentational, properties are *not* required for phenomenal consciousness.

[22] Also there are those who acknowledge phenomenal concepts but who reject the phenomenal concept strategy and physicalism itself (Chalmers, 2006).

In contrast, HOR theorists contend that a mental state M is phenomenally conscious if and only if it is appropriately represented by a higher-order state M*.[23] For example, HOR theorists claim that my being in a phenomenally conscious state of seeing the blue sky can be explained in terms of my undergoing an appropriate representation of my state that represents the blue sky. Such a representation is *higher-order*, or *metarepresentational* because it represents (or is about) another mental state—in this case, it represents my own visual state of seeing the blue sky.

The aim of this dissertation is to develop a novel intrinsic higher-order account. I will not be concerned with the debate between FOR and HOR theories. While FOR theories have many virtues, they suffer from one fundamental vice: existing FOR theories fail to meet the *conjunction* of conditions D1 and D2. They cannot draw the conscious/nonconscious distinction *and* explain the distinctive features of phenomenally conscious states.[24]

Consider Tye's "PANIC" theory, according to which phenomenally conscious states are constituted by poised, abstract, non-conceptual, intentional contents. There are quite obviously abstract, non-conceptual, intentional contents in cases of non-conscious representation, and adding "poisedness" does not solve the problem. To be poised is just to be available to inform beliefs and desires. Whether or not one rejects non-conceptual

[23] There is some disagreement about whether the first-order object must be an actual state of which the subject is undergoing. I will return to this issue Chapter 2. Moreover, to the best of my knowledge all HOR theories also set constraints on the first-order object of the higher-order state, whether that object is actual, or merely an intentional inexistent. The constraints that I set will be outlined in Chapter 2 section 3.2.

[24] Carruthers (2005) and Kriegel (2009) argue similarly.

content doesn't really matter. Even if the conceptualist substitutes conceptual for non-conceptual (a "PACIC" theory, if you will), there is nothing about first-order conceptual content that will suffice to make the state conscious. I will not argue further for this claim. I will simply assume it for the sake of the arguments that are the main focus: arguments against existing HOR theories and in favor of a novel version.

I propose to assume, also, the failure of several standard objections to representationalism in general, regardless of the FOR/HOR distinction. For example, (i) the possibility of non-representational experiences (Block, 1996; 2003), (ii) counterexamples that are supposed to show that one can undergo states that have the same representational content, but different phenomenal character (Peacocke, 1983; Block, 1996; Lopes, 2000), (iii) counterexamples designed to show that one can undergo a state that has the same phenomenal character, but different representational character (Block, 1990). Several authors have already put forth plausible responses to these proposed counterexamples, and for the purposes of this dissertation, I will assume that those responses show that the objections to representationalism in general are at least not decisive.[25] Furthermore, my goal is not to defend representationalism in general from the ground up, but rather to increase the plausibility of one specific branch of representationalism, *viz.* the higher-order branch. I will therefore assume that representationalism of one kind or another is currently a possible path of explanation that is worth pursuing.

[25] Regarding (i) see Tye (1995, 1995b), Lycan (1996), Carruthers (2000), and Kriegel (2009). Regarding (ii) see Harman (1990), DeBellis (1991), Tye, (1992, 1996, 2000: Ch. 4), Carruthers, (2000, p. 117), Dretske (2000), and Kriegel, (2009). Regarding (iii) see Tye (2000), Shoemaker (1994a, 1994b, 2002), and Kriegel (2009).

Among HOR theories themselves, one may draw a further distinction. On the one hand, there are those HOR theories that take the higher-order component to be a numerically distinct state. Following Gennaro (2006, 2012), I call these *extrinsic* HOR theories (or EHOR theories). On the other hand, there are those HOR theories that take the higher-order element and its first-order target to be parts of the same state. I call these *intrinsic* higher-order theories (or IHOR theories), because they hold that consciousness is an intrinsic property of a mental state.

IHOR theories are typically referred to as "self-representational" theories. However, drawing the distinction between HOR theories and self-representational theories as a distinction between "orders" is mistaken, or misleading at best. It suggests that the main difference between the two kinds of theory is that one is higher-order (that is, it includes a metarepresentational commitment) and that the other is "same-order" (that it does not include a metarepresentational commitment).[26] But, *both* extrinsic higher-order theorists and self-representationalists are committed to some kind of metarepresentationalism. The characteristic which distinguishes the two kinds of view is whether the metarepresentational component to which they appeal is numerically distinct from, and extrinsic to, the conscious state, as extrinsic theorists maintain, or whether it is intrinsic to the conscious state itself, as "self-representationalists" maintain. For this reason I take the four representative versions of self-representational theory discussed

[26] Kriegel (2006) introduced 'same-order', into the literature, but has since abandoned it.

below to be members of a subset of higher-order theory,[27] and, at the risk of adding to the terminological quagmire of contemporary philosophy of mind, I will use the extrinsic/intrinsic higher-order dichotomy rather than the higher-order/self-representational dichotomy.[28] The mental quotation account that I develop in Chapter 4 is an intrinsic higher-order thought theory.

HOR theorists typically argue that their view presents an advancement over FOR theories. In Chapter 2 I argue that, in spite of the explanatory advancement, EHOR theories face substantial challenges, and that these motivate IHOR theories. Perhaps the main shortcoming of EHOR theory is the problem of higher-order misrepresentation, which will be developed in detail in Chapter 2 and revisited in Chapter 3. The problem arises from the assertion that the higher-order element is part of a state that is numerically distinct from its first-order target, and this lack of "intimacy" allows for the possibility of mismatching first-order and higher-order states, which extrinsic views cannot adequately accommodate.[29]

Others before me have argued that higher-order misrepresentation motivates intrinsic HOR theory. I agree. However, in Chapter 3 I argue that existing IHOR theories

---

[27] Notice that three of the four self-representational theories were actually introduced as variants of higher-order theory, *e.g.,* Gennaro's WIV, Van Gulick's HOGS model, and Carruthers' dual-content theory.

[28] Gennaro (2006) has already introduced similar terminology (pure self-referentialism v. extrinsic HOR theory), but the terminology hasn't caught on. For example, Block (2011) continues to call self-representationalist views "same-order" or "same-state" views, but the difference between the two is not merely terminological. It marks a substantive distinction that has explanatory consequences, or at least, I will argue for such consequences in Chapters 2 and 3.

[29] This is the further component of the intimacy problem mentioned in section 2. It will be explained in more detail in Chapter 2.

continue to encounter difficulties addressing higher-order misrepresentation, and in Chapter 4, I propose an alternative intrinsic account that is immune to higher-order misrepresentation and which has something to say about the initially puzzling data introduced above. This is "the mental quotation model of consciousness," according to which the higher-order element of a conscious state is a quotational thought that takes sensory contents as its object. Mental quotations are representations that target (or quote) other states, but their mode of representation is more intimate than the relations posited by either EHOR theories or existing intrinsic views. On the quotational account, a higher-order state latches onto, or envelops, its first-order target, integrating that sensory content into itself.

With this background in place, in the next chapter I will discuss the main tenets of extrinsic higher-order theory as well as some main objections to the view.

Chapter 2: Extrinsic HOR theory: Higher-order Relevance, Higher-Order Misrepresentation, and Fineness of Grain

This chapter reviews and assesses traditional higher-order theory as developed and defended by David Armstrong (1968, 1984), William Lycan (1987, 1996), David Rosenthal (2005) and others. Following Gennaro (2006), I will call such views *extrinsic* higher-order theories (EHOR theories for short). In section 1, I present the main commitments of EHOR theory. In Section 2, I illustrate how the view handles the puzzling features that were introduced in Chapter 1. In section 3, I review two problems for EHOR theories concerning the relevance of higher-order representation. These are the "rock problem" (Goldman, 1993; Dretske, 1995; Stubenberg, 1998), and the "too-easy problem" (Rey, 2008). I will argue that neither of these objections are fatal problems for the view. In section 4 I review the problem of higher-order misrepresentation as the problem is typically presented (Neander, 1997; Levine, 2000; Kriegel, 2009; Block, 2011). I then go on to refine the problem, presenting an argument from the apparent fineness-of-grain of experience. I also review some recent empirical data that are thought to bear on the issue of higher-order misrepresentation (Lau and Rosenthal, 2011; Lau and Brown, forthcoming). I argue that the problem of higher-order misrepresentation, when construed properly, is a pivotal objection to traditional EHOR theories, and that it motivates major revisions to the higher-order framework. In particular, I will argue that it motivates a self-representational (*intrinsic* higher-order) theory.[1,2]

[1] As I will make clear in Chapter 3, the relevant version of contemporary self-representational theory is a version of higher-order theory. It is not a same-order theory,

1. Traditional Extrinsic Higher-Order Theory

As discussed in Chapter 1, FOR theorists contend phenomenal consciousness consists in a subject S undergoing a mental state M that represents appropriately. In contrast, EHOR theorists contend that phenomenal consciousness consists in S undergoing a higher-order state M*, which represents a mental state M appropriately. Some EHOR do not require an *actual* mental state M. Some EHOR theorists (*e.g.,* Lycan and Carruthers pre-2000), on the other hand, specify that the state M must not only be actual but must also represent appropriately.

As an example, EHOR theorists claim that my being in a phenomenally conscious state of seeing the blue sky can be explained in terms of my having an appropriate representation that I am in a state of seeing the blue sky (and, perhaps, an appropriate blue sky-representing state too). Such a representation is higher-order, or metarepresentational, because it represents (or is about) another mental state—in this case, it represents my own visual state of seeing the blue sky.

In addition to differing on the necessity of there being a first-order state, and what its properties might be, theorists also differ on what appropriate higher-order representation consists in. For example, according to higher-order perception theorists (HOP theorists) such as Lycan (1987, 1996), an appropriate higher-order representation is perception-like.

---

as Kriegel (2006) originally characterized it, and as at least one theorist still does (Block, 2011). I *think* Kriegel himself has finally come around to this idea.

[2] Gennaro (1997, 2012); Carruthers, (2000, 2005); Van Gulick, (2004, 2006); and Kriegel, (2006, 2009) are some others who have argued that higher-order misrepresentation motivates a self-representational view. In section 4 I present my own novel argument for that same conclusion.

On Lycan's view, there is literally an internal sense organ that scans the outputs of first-order sensory systems, and which generates a higher-order *perceptual* state. Many authors have argued that the prospects for a faculty of inner sense look increasingly grim (Carruthers, 2000, 2005, 2011; Kriegel, 2009). I will set inner-sense theories aside and focus my attention on EHOT theory. According to higher-order thought theorists (HOT theorists), the higher-order state is a *thought* about its first-order target. A subset of HOT theory, actualist HOT theory, describes a subject as being in a phenomenally conscious state M if the subject undergoes an appropriate HOT M* about M (Rosenthal, 2005; Lau, 2008; Weisberg, forthcoming, 2011). In contrast to the actualists, in the dispositionalist version of HOT theory, a mental state M is conscious when it is *available* to the relevant higher-order thought-producing faculty that would generate a higher-order state M* that represents M (Carruthers, 1996; 2000, 2005).[3]

Originally EHOR theory was characterized as a kind of monitoring. On this construal of the view, a higher-order state (or some mechanism responsible for producing such states) monitors the outputs of first-order sensory systems. For example, a first-order/world-directed mental state M, when monitored, occurs in rough simultaneity with a higher-order state M*, which renders the state M phenomenally conscious. This is roughly how Armstrong (1968, 1984) characterized the view and to some extent how Lycan still characterizes the view (1987, 1996). Also (while Rosenthal hadn't yet introduced the distinction between *creature* and *state* consciousness), the view was

[3] In the next chapter we will see that, while Carruthers introduced his account as a dispositionalist HOT theory, it is actually a version of self-representational (HOR) theory (Carruthers, 2011). As I argue, the leading versions of self-representational theory are a subset of higher-order theories.

implicitly intended to provide a naturalistic explanation of what makes a mental *state* conscious.[4]

Combing the monitoring characterization of EHOR theory with the endorsement of phenomenal consciousness as a species of mental state consciousness, it is plausible to expect all EHOR theorists to think that the existence of an actual first-order state is one of the necessary conditions for consciousness, *i.e.,* that a subject S is in a conscious state when, and only when, S undergoes a HOR that represents an actual (and appropriate) first-order state that S is also undergoing. Call this requirement the Existence Condition (EC).

> (EC:) A subject S undergoes a phenomenally conscious state if and only if S undergoes an *actual existing* first-order state that is represented by a higher-order state.

Some EHOR theorists actually reject the existence condition. In particular, the most active defenders of EHOT theory (David Rosenthal and Josh Weisberg) argue that a HOT itself is sufficient for phenomenal consciousness, and that a HOT need not target an actual existing first-order state. For example, Rosenthal has said "…HOTs determine what it's like for one to be in various conscious qualitative states. So erroneous HOTs will in this case result in there being something it's like for one to be in a state that one is actually not in" (2005, 209). And,

> If one has a sensation of red and a distinct HOT that one has a sensation of green, the sensation of red may nonetheless be detectable by various priming effects. But what it will be like for one is that one has a sensation of green. *Similarly if one has that HOT with no relevant sensation at all* (2009, 249, my emphasis).

[4] In Chapter 1 I already discussed that, on the vast majority of views, phenomenal consciousness is presumed to be a kind of mental state consciousness.

And, "Conscious states are states we are conscious of ourselves as being in, *whether or not we are actually in them"* (2002, 415, my emphasis).

Similarly, Weisberg (2010) argues that what is conscious is always an intentional object. In veridical cases of higher-order representation the intentional object happens to correspond to an existing mental state. But there are also erroneous cases of higher-order representation, including erroneous cases wherein there is no actual first-order state at all. Those latter cases involve intentional *in*existents, and according to Rosenthal and Weisberg, when they are the objects of a HOT, they too suffice to underwrite mental state consciousness. In targetless HOT cases, the conscious state is a state that does not actually exist.[5] The distinction between those versions of EHOR that endorse EC (existence versions) and those that reject it (inexistence versions) will be important when assessing whether or not EHOR has the ability to address the problem of higher-order misrepresentation.

Thus, there are many different kinds of EHOR theories, and between some of these there are considerable differences. Presently, I want to abstract away from most of these differences to focus on the main commitments that all EHOR theorists share, including both existence versions and inexistence versions of the view.

[5] Targetless higher-order representations will be discussed in more detail in section 4. Another possibility is to construe EHOR theory as a kind of creature consciousness, rather than state consciousness. Brown (Lau and Brown 2011) hints at such a move, but does not develop the idea much. I will return to this issue in section 4.3.

The first commitment is to what Rosenthal calls the "transitivity principle," which asserts that conscious states are states a subject is *aware of* being in.[6] There are different ways to motivate the transitivity principle. Rosenthal (2000, 2002, 2005) maintains that the transitivity principle is the best way to capture our commonsense understanding of mental state consciousness. Gennaro argues that it's a conceptual truth (2012, 28). Lycan (2001) maintains that the principle is a stipulative definition that corresponds to at least one commonsense notion of 'consciousness'. He then puts it to work in an argument for higher-order theory.

Lycan argues that,

(1) A conscious state is a mental state whose subject is aware of being in it.

(2) The 'of' in (1) is the 'of' of intentionality; what one is aware of is an intentional object of awareness.

(3) Intentionality is representation; a state has a thing as its intentional object only if it represents that thing.

Therefore,

(4) Awareness of a mental state is a representation of that state.

And therefore,

(5) A conscious state is a state that is itself represented by another of the subject's mental states (2001, pp. 3-4).

The simple argument suffers from at least two flaws that are relevant to my main concerns: i) it begs the question against FOR theories, and ii) as an argument for EHOR theory exclusively, it is invalid.

[6] Of course FOR theorists deny the transitivity principle. They maintain that conscious states are essentially states that make us aware of things in the world, excluding our own mental states, *i.e.,* conscious states need not make us aware of our own states.

Premise 1 asserts the transitivity principle as a stipulative definition, and it straightforwardly begs the question against first-order theorists. FOR theorists deny that conscious states are states a subject is aware of being in. They claim that conscious states are states that make one aware of whatever those states represent, but conscious states need not represent other mental states. So in the absence of further argument, (1) will be unacceptable to FOR theorists. Moreover, (5) simply does not follow. It does not follow from (2) and (3) that a conscious state is a state that is itself represented *by another of the subject's* mental states. It only follows (granting the truth of premise 1) that a conscious state is a state that is itself represented. However, as we will see in the next chapter, intrinsic theorists argue that when a mental state is conscious, the relevant first-order state and the higher-order state that represents it might not be realized by numerically distinct mental states. Rather, they might be realized by different components of the same complex (or global) higher-order state, and this is consistent with the truth of 1-4. Thus, 1-4 could be true, while 5 is false. In spite of these two weaknesses, Lycan's argument does capture an insight of HOR theory, and the motivation for the transitivity principle and its metarepresentational component can be recast differently, in non-question-begging terms.

Consider again the distinction between conscious and nonconscious perception that was discussed in Chapter 1. One example referred to came from the two visual systems hypothesis, according to which percepts along the dorsal stream have no corresponding *conscious* awareness. Consider my non-conscious apple-on-the-desk perception along the dorsal path. In such a state I would not consciously experience anything, even though the relvant state can figure in the casual path leading up to several different behaviors, *e.g.,*

reaching out to grab the apple when overtaken by hunger. But if ultimately this process is to be characterized computationally, there are relevant computational relations that would contribute to the production of such actions. There must be an apple-on-the-desk representation, because there must be some representation to compute (think Fodor, 1975). The point is that perception and/or awareness, whether phenomenally conscious or not, requires some kind of representation (among other things). But now, one might plausibly think that to be aware of these perceptions (in an as of yet unspecified sense of 'aware'), they too would need to be represented. For example, my apple-on-the-desk perception itself must be represented. And now the critical question is whether or not a detailed description of that kind of state, *i.e.,* a detailed description of *some kind* of higher-order representation (whether it is a thought or a perception, an actual representation or a disposition to form a HOR), is sufficient to explain the distinctive features of phenomenal consciousness. FOR theorists, among others, argue that it is not sufficient. HOR theorists argue that it is. The ultimate success of this argument for HOR theory depends on the further details of the view. However, as a general argument for HOR theory, *i.e.* for a theory that requires some kind of metarepresentational component, it does not beg the question against FOR theorists straightaway, as Lycan's does by stipulating in the first premise of his argument that consciousness is higher-order. Nor does it rely too heavily on the common sense notion of consciousness, nor does it declare the transitivity principle as a conceptual truth. The above might not be completely convincing or *prove* HOR theory, but I think it gestures towards a general argument for the HOR framework that I assume can be made to work.

A second main commitment of EHOR theorists (all HOR theorists for that matter), which may be held with varying degrees of strength, is the commitment to what Neander (1997) calls the "division of phenomenal labor," according to which first-order representational properties determine and explain qualitative character (how a mental state M represents anything in the first place), and higher-order representational properties determine and explain phenomenal character (the what it's likeness of M's subject being in M). One reason that the division can be held with varying degrees of strength has to do with whether or not one accepts or rejects the existence condition. A strong endorsement of the division would require that there be an actual first-order state to determine the qualitative properties of the phenomenal state. On Rosenthal's and Weisberg's view, an actual state M is not required. Their view can be seen as involving a weaker endorsement of the division of phenomenal labor. Since there is no actual first-order state to determine the qualitative properties of the state, they must be determined by the higher-order state itself, *i.e.,* the qualitative properties of the phenomenal state are somehow determined by the first-order state *as it is represented by* the higher-order state, even though that first-order state could be an intentional inexistent.

Whether one holds the division weakly or strongly, it is important to notice that the higher-order properties are supposed to determine the phenomenal character of a numerically distinct, extrinsic state, *viz.,* the state M, whether it exists or not. Distinctness, then, is the third main commitment of traditional EHOR theorists: the conscious states

that we are aware of being in are numerically distinct, extrinsic states from their higher-order consciousness issuing counterparts.[7]

2. EHOR and the Puzzling Data

The ultimate explanatory success of EHOR depends on the details of the specific version in question. These details will be the focus of the section. I quickly review how EHOR theory handles the puzzles D1-D5, discussed in Chapter 1. I do not intend this as a full-throated defense of HOR theory. My goal in this dissertation is to argue that if some HOR theory is indeed correct, then IHOR theory is superior to EHOR theory. Furthermore, the mental quotation version of IHOR to be introduced in Chapter 4 is the best of version the latter. While IHOR is ultimately the better view, I think (as do other IHOR theorists) that EHOR has many virtues that are worth preserving.

D1 What it's Like

To meet the essential adequacy condition and explain the subjective feel of phenomenally conscious mental states, EHOR theories can appeal to the dual contents of the relevant first-order/higher-order pair. In virtue of having fine-grained/world-

[7] There is also one major choice-point that will divide HOR theories overall, *i.e.,* both extrinsic theories and intrinsic HOR theories. This is whether or not to construe the higher-order component as an explicit representation or as an implicit representation. Since all extrinsic theorists construe the higher-order representation as explicit, and only intrinsic theorists construe it as implicit, I will postpone further discussion until the next chapter. But notice that the explicit-implicit distinction does not itself mark the division between extrinsic and intrinsic theories. That the explicit-implicit distinction overlaps with the extrinsic-intrinsic distinction is, I think, incidental.

representing first-order contents, such states will have qualitative character.[8] For example, a first-order visual state representing a tree will partly determine the worldly properties represented in experience, and representing such features will allow the creature to navigate the world (*e.g.,* it will be able to avoid the tree, rather than bump into it, climb the tree to pick fruit, &c.). This much can be explained by a first-order theory, but it doesn't tell us what *the experience* of the tree is like. A higher-order theorist can explain that by appeal to the higher-order, experience-representing content of the higher-order state. In virtue of undergoing a representation of that first-order worldly state as a state the subject is undergoing, the experience of seeing the tree will be like something for the subject who undergoes it. There will be something it's like to perceive the many leaves fluttering in the wind, the different shades of green in spring, and the darkness or lightness of the bark.

D2 The Conscious/Non-Conscious Distinction

Another virtue of the traditional EHOR framework is that it provides the foundation for a plausible explanation of the distinction between qualitative/sensory character and phenomenal. More specifically, EHOR theorists can explain that, *e.g.,* the non-conscious first-order visual percepts along the dorsal path (or a nonconscious visual percept of objects in cases of unilateral neglect) are genuine qualitative states (they represent, among other things, various paradigmatic qualitative features of objects, *e.g.,* shape, color, size, &c.), but since there is no corresponding conscious awareness, there is nothing it is like for the subject who undergoes such states. Since phenomenal character

[8] 'World-representing' here does not imply that one must endorse an externalist theory of first-order mental content.

is the what-it's-like component, such states lack it and are not phenomenally conscious. The reason why those states are not conscious is that they are not appropriately represented by a higher-order state.

Notice too that this not only gives a straightforward way to draw the conscious/nonconscious distinction, but also it does so *while explaining* the distinctive subjective feel of phenomenally conscious states, and as we will see, while also explaining some of the other core features of phenomenally conscious states. Thus, HOR theories can meet the conjunction of D2 and D1. As argued in Chapter 1, meeting the conjunction D2 *and* D1 was the main pitfall of FOR theory.

D3 Intimacy

Recall from Chapter 1 that phenomenally conscious experiences seem to involve an awareness that is "intimate," but that there are at least two distinct notions of intimacy in the literature: a temporal notion (immediacy) and an epistemic notion (noninferential). I already discussed how FOR theorists might handle intimacy/immediacy claims. Despite the fact that many have argued that EHOR theory, in particular, encounters problems handling such claims—because of the "distance" between the first-order and higher-order states (Goldman, 1993; Natsoulas 1993; Moran, 2001; Kriegel, 2006, 2009)—EHOR theories can handle most intimacy claims in much the same way as FOR theories.

Here is how Kriegel characterizes one of the problems for EHOR theories.

> Suppose S has a conscious perception of a tree. According to [extrinsic] higher-order representationalism, the perception, M, is conscious because S has another mental state, M*, which is an appropriate higher-order representation of M. Now, surely *M* normally has a role in the causal process leading up to the formation of M*. Just as the tree normally has a central role in the causal process leading up to the perception of it, so the perception itself normally has a central role leading up to the higher-order representation of it. Arguably, M* would not *be* a representation of M if that were not

> the case. This means that the formation of M* is not exactly simultaneous with the formation of M. Rather, there is some sort of (temporally extended) causal process starting with M and ending in the formation of M*. (2009, 152, fn. 73).

It should be clear from the above passage that Kriegel is discussing the temporal sense of 'immediacy'. The problem for EHOR theories is supposed to arise from the claim that the formation of M* and M are "not exactly simultaneous," or that they are severed by "some sort of (temporally extended) causal process." That is supposed to interfere with our phenomenally conscious states *seeming* immediate, because, if the process leading from M to M* is mediated by a temporally extended process, then our awareness of such states should not seem immediate. But again, this involves the assumption that any (relevant) temporally extended process in the brain should seem so in experience. This assumption is unwarranted.

As I argued in Chapter 1, this is actually not a problem specifically for EHOR theorists. If it is a problem at all, it is a problem for any account which posits a temporally extended process between a state realizing an object of representation and a state realizing our awareness of that object. If any temporally extended process rules out immediacy, then FOR theorists too will encounter difficulties (*e.g.,* the process leading from an ANIC state to a PANIC state is temporally extended). But it is difficult to see exactly why any temporally extended process should matter in the first place. Again, as argued in Chapter 1, one might challenge the temporal notion of immediacy altogether on the basis of the distinction between how long a process actually takes and how time is represented in the mind (Dennet and Kinsbourne, 1995a). What might actually be a temporally extended process could very well *seem* immediate.

One common way of capturing the difference between the immediacy of conscious states and those states that do not seem immediate is to claim that conscious states are generated non-inferentially. This line of thought is developed by Rosenthal (1990, 1993, 2005). Rosenthal has phrased his response to immediacy differently over the years, but I think at this stage the following is a fair characterization. On Rosenthal's view, conscious states do indeed have the appearance of immediacy. However, to seem immediate, conscious states don't necessarily have to be unmediated (by inferences). They simply have to result from intervening processes that may in fact be mediated (by inferences), but these processes are not themselves conscious. As Rosenthal argues, immediacy may be explicable in terms of the *seeming* immediacy with which a higher-order state makes its subject aware of its first-order target. As long as the process from a first-order target to its representation by a HOT is not *consciously* inferential, it will not seem to the subject to be mediated. Rosenthal's reply is persuasive. We would need further explanation to conclude that *any* mental process that is mediated *must seem* mediated.

However, Kobes (1995) raises an additional problem for Rosenthal's account of immediacy. Kobes argues that Rosenthal's account requires modification. For the requirement that a HOT merely be assertoric is too weak, because we can have genuinely assertoric HOTs generated non-inferentially that target *other people's* first-order states, but we wouldn't want to say that those states (in other people!) become conscious as a result. Kobe argues for a telic, or desire/action-like force, rather than mere assertion. I will discuss the issue raised by Kobes in more detail in section 3. Here I just want to demonstrate that there are possibilities for the EHOR theorist to handle immediacy.

D4 Phenomenal Knowledge

Like FOR theorists, EHOR theorists too might be able to meet D4 insofar as they can appeal to the phenomenal concept strategy, but this might not have much to do with the first-orderness or higher-orderness of the respective theories.

Recall that, as originally intended, the phenomenal concept strategy is supposed to be available to any physicalist theory of phenomenal consciousness. If so, then HOR theorists can help themselves to the strategy, and thereby, assuming the strategy is successful, meet D4. Also recall, though, that in contrast, Caruthers (2005) argues that a certain kind of HOR theory is *required* to explain how we can acquire purely recognitional concepts, his preferred view of phenomenal concepts, and if that is true, then the phenomenal concept strategy might not be available to all EHOR theorists, for it would be available only if such theorists could provide or were consistent with an account of how we acquire purely recognitional concepts. Again, much depends on whether phenomenal concepts are purely recognitional in the way that Carruthers says they must be. If phenomenal concepts are purely recognitional in Carruthers' sense and EHOR theories can explain how we could acquire such concepts, then they would have a way of meeting D4. If phenomenal concepts are not purely recognitional in Carruthers' sense, then EHOR theorists may appeal to some other account of phenomenal concepts to meet D4. Another possibility is to abandon the phenomenal concept strategy altogether and meet D4 in some other way, à la Tye (see Ch. 1 fn. 23).

As briefly sketched in this section, EHOR has many virtues. It has solid accounts of D1-D4 and is generally well motivated by the more general argument given in section 1. However, there are further difficulties the view encounters that arise particularly from

positing a numerically distinct metarepresentational state. I deal with those difficulties in the next two sections.

3. Two Relevance Problems for EHOR

In this, and the next, section, I discuss the higher-order issues that fall under D5 from Chapter 1. The two problems I will focus on in this section are the *rock problem* and the *too-easy problem*. Both of these problems challenge EHOR theorists to explain the relevance a higher-order state has to making its target state phenomenally consciousness. I argue that the two problems ought to rule out certain versions of the view, but that neither presents a knock-down reason for doubting the plausibility of HOR theory in general. The third problem (higher-order misrepresentation) has generated a lot of discussion over the past few years, so I will devote the entirety of section 4 to it. Again, here I am not trying to establish the *truth* of EHOR theory. More could be said in its defense, but that is not my aim. In this chapter, I am arguing that HOR theories are plausible enough to consider seriously. The ultimate goal of this dissertation, though, is to establish the best HOR theory, so most of the subsequent argument assumes the truth of HOR theory.

3.1 The Rock Problem

EHOR theorists hold that a subject S undergoes a conscious state M if and only if S undergoes an appropriate HOR M*, which represents M. We tend to think, though (with good reason), that a rock does not become conscious when S undergoes an appropriate mental representation of it. But why, then, should S's own mental state become conscious

when S undergoes a mental representation of it, as EHOR theorists maintain M does? Proponents of the rock objection argue that EHOR theorists have no principled way of answering the question, and consequently, that they are left committed to the implausible conclusion that all kinds of object (both animate and inanimate) can be rendered conscious in virtue of being mentally represented by someone's, or some organism's, mental state (Goldman, 1993; Dretske, 1995; and Stubenberg, 1998; Kriegel, 2009).[9]

The standard reply is that perceptions of rocks are mental states, but rocks themselves are not. By hypothesis, phenomenal consciousness is supposed to be a property of mental states, so there is no reason to expect that a mental representation of a rock would render the represented rock conscious. We only consider mental states to be the kinds of things that can be conscious. As Lycan puts it, "What is it that is so special about physical states of that certain sort, that consciousness of them makes them conscious? That they are themselves mental… It seems psychological states are called 'conscious' states when we are conscious of them, but nonpsychological things are not" (1990/1997, 758-759).

Lycan's reply sounds like another stipulation, but a stipulation doesn't tell why it only psychological states can be conscious. There is more that higher-order theorists should say to rule out rocks as candidates for mental state consciousness – and there is more they *can* say.

For example, it is not simply that we merely stipulate that only psychological states conscious, but rather, there is no independent reason in support of a theory that would have us attribute consciousness to rocks and other non-psychological entities on the basis

[9] Sometimes the rock problem is called the "problem of generality." For example, see Van Gulick (2006).

of their being mentally represented. There is, however, independent reason in support of the competing claim that phenomenal consciousness requires, and is parasitic upon, first-order qualitative character. The division of phenomenal labor requires that there must be an appropriate FOR (even if it is a non existent FOR that is merely represented by a HOR). A rock is not an appropriate FOR. Entities such as rocks and other nonrepresentational states (including certain kinds of internal states) are just that: they are not representational. Moreover, some representational mental states themselves are not *appropriately* representational, and we should not be tempted to attribute consciousness even to them.

My own understanding of an "appropriate" first-order state is just the output of something like a prototypical sensory modality. Picciuto and Carruthers (2011) suggest constraints that a constitutive of a prototypical sensory modality. Here I assume these constraints. A prototypical sensory modality will: (1) be sensitive to some range of physical energy or set of physical properties, (2) include a detector mechanism that transduces that energy or those properties into informational signals sent to the central nervous system where (3) they are used to guide the intentional behavior of the organism. In addition, a prototypical sense will (4) have as its evolutionary function the detection and representation of the physical energy or properties in question, and (5) will issue in non-conceptual representations with mind-to-world direction of fit.[10] While a full account of a prototypical sensory system would no doubt need to include some specification of

[10] It's not clear how much 4 really matters, but again, this is the outline of a prototype, and our prototypes are the sensory systems of organisms evolving in the natural world. That by no means suggests that sensory systems are limited to "natural" organsims.

the comparative importance of each component relative to the others, a simple listing of the five components will be sufficient for my purposes here.

Rocks and inanimate objects clearly do not the fit bill, and can be plausibly ruled out. For if rocks are not representational at all, then they cannot be appropriately representational. Thus, there is no need merely to stipulate that only psychological states can be conscious. There are independent reasons that are consonant with broader models of explanation for why a conscious state must be psychological.

While rocks and other inanimate objects can be safely ruled out, there is another problem that emerges from the rock problem. One might wonder why representing another person's appropriate first-order mental state does not render it conscious. The potential problem is that EHOR theory might be committed to the puzzling claim that a HOR in one person can render some other person's FOR conscious. Call this the "other 'other minds' problem."

Refining the Rock Problem: The other other-minds problem

If consciousness amounts to having an appropriate HOR of an appropriate FOR, it seems possible to have a HOR that targets *another* person's appropriate FOR. But we tend to think that doing so would not render the other person's appropriate FOR conscious. In the absence of further explanation, EHOR theory would be left with the counterintuitive conclusion that a HOR in one person could render a first-order state in another person conscious.[11]

[11] Kobes (1995) touches on this point as a problem for Rosenthal's EHOT theory, but I think the problem can be developed even more as a problem for EHOR theory in general.

Here too, the EHOR theorist might try a stipulation: targeted states are conscious only in virtue of being targeted by a HOR in the same subject, which best captures our common sense notion of consciousness. But this does not add very much to the *explanation* for why targeted states in others are not conscious. It suffers from the same problem as Lycan's initial response to the rock problem: stipulations aren't explanations.

There is more that can be said, though, beyond merely stipulating that for consciousness to obtain the relevant FOR/HOR pair must be intrasubjective. One might be tempted to argue that a FOR in one person cannot *directly* cause a HOR in another person, as they typically do intrasubjectively when issuing in consciousness. This is not the most promising defense. The challenge is to provide an adequate analysis of 'direct'. While representational theories are *representational* theories, representationalists believe that representational states are typically realized by brain states. Thus, the relationship between brain states may be relevant. One's own brain state cannot be directly caused by another's in the same way. The difficulty is to satisfyingly account for the relevant "way." One obvious possibility is to appeal to neurophysiological properties, but that would extend beyond the constraints of a *representational* theory.

Another problem with the claim that the states in other minds are not directly caused is that, while EHOR theorists claim first-order states sometimes (or even usually) cause their targeting HORs, they will surely agree that a HOR need not *always* be caused by an appropriate FOR. So the question that arises is: Why should it matter that the relevant HOR is not directly caused in other-minds cases?

More plausibly, recall Rosenthal's original claim that conscious states are reduced to states we represent ourselves as being in. One might then argue that only one's own

states are represented *as one's own, i.e.,* phenomenal consciousness just is the way experiences seem *to the subject.* James' appropriate FOR will not seem like anything for Sally when represented by James, though James might be in a state that seems like something to James (even if part of that seeming is that James is in Sally's appropriate FOR). Notice this line of reasoning is available to both existence versions and inexistence versions of EHOR. On both views the conscious object is an intentional state. Even if I represent myself as being in someone else's FOR, that FOR is an intentional object represented by *my* HOR.

To sum up, the rock problem is originally intended to challenge the relevance higher-order states have to consciousness. But, as we have seen, the rock problem actually winds up emphasizing the importance of there being an appropriate HOR, which depends, in part, on whether or not it represents the subject as being in an appropriate FOR. But this is something that EHOR theorists already acknowledge, so the rock problem is not quite the problem that many have thought it to be.

3.2 The "Too-Easy" Problem

A related objection is Rey's (2008) "too easy" objection. The objection is related to the rock objection because, like it, the too easy objection urges one to reflect on the explanatory relevance of higher-order representations. Rey does not intend his objection to apply to EHOR theory in general. Rather he intends it to pose problems for EHOT theories in particular. This is because:

> higher-order sensing or perceptual theories… seem to involve presumptions about processes in addition to specific intentional contents. At least for purposes here I am prepared to suppose that these further processes may be sufficiently rich to supply the conditions –the right "connection"– to render a state conscious, along

> the lines of, say, a causal theory of perception in general, whereby someone sees something only if that thing caused a visual experience of it in the right sort of way (2008, 9).

One thing to notice is that HOT theories too may involve "presumptions about processes in addition to specific intentional content," so while Rey poses the problem for EHOT theories in particular, I will discuss it as a problem for EHOR theories in general. The main point of the too-easy objection is that there must be more to phenomenal consciousness than higher-order representation, because higher-order representation is "too easy." Higher-order representations seem present in at least three cases about which we would be disinclined to say that the subject is conscious. According to Rey, HORs seem to occur in ordinary computers, "intra-modular" HORs seem possible in humans, and non-conscious HORs seem possible when considering intuitive "Freudian" cases of the unconscious. I will explain each of these components of the too easy objection in turn and explore possible replies on behalf of the EHOR theorist.

HORs in Computers

One component of the too easy objection arises from the existence of HORs in ordinary computers. The problem is essentially this. Ordinary computers undergo higher-order states. But it is intuitively implausible to attribute phenomenal consciousness to such a state in, say, a laptop. A state S2 in one's laptop might represent that it is in S1, but intuitively, that wouldn't make S1 phenomenally conscious.

The main option here is to reject that ordinary computers undergo appropriate HORs, because they do not undergo appropriate FORs. A computer must first be capable of perception before it can undergo *conscious* perception, and "ordinary" computers, such as laptops, surely do not have sensory modalities. If they lack sensory modalities they

lack appropriate FORs. As discussed in the previous section, the outputs of a prototypical sensory modality have distinctive features most significantly they are fine-grained, nonconceptual states with mind-to-world direction of fit. Since laptops have no sensory modalities (and no other way of producing fine-grained, analog representations with mind-to-world direction of fit), none of the HORs that ordinary laptops can be plausibly thought to undergo are appropriate. Notice that hooking up a camera and/or microphone will not suffice, because there would still be no good reason to think that such a system undergoes representational states that meet condition 5.[12]

The constraint of having an appropriate FOR rules out ordinary computers (including existing supercomputers). However, it does not rule out the possibility of a more sophisticated computer, *e.g.,* some kind of sentient robot, being conscious—the kind of machine that meets the above criteria, transducing information from the environment and using it to guide decision making and action planning, moving about in the world appropriately, and importantly, doing so while also having appropriate HORs of its appropriate FORs.

What this amounts to is the assumption that a system with the right kind of representational states and the right kind of cognitive architecture will realize conscious states. But there is nothing startling about that claim. For that is just what the (computational) representationalist claims anyway. The difference here is that the kind of system that realizes that overall sensory and cognitive architecture is not much like an ordinary laptop at all. In fact, there is a range of cases between "ordinary laptop" and "sentient robot" to consider, but I am going to set those intervening cases aside and

[12] I will discuss this issue in more detail in Chapter 5.

simply concede that it is *possible* to construct an adequately sophisticated computing system that would be conscious, *if* it had the right cognitive architecture and it had something sufficiently resembling a prototypical sensory modality (as specified above). For if it had the right cognitive architecture and something resembling a prototypical sense, there would be little motivation left for denying that such an entity is conscious, perhaps with the exception of our folk intuitions about the kinds of organisms that could possibly be conscious. These intuitions might, in fact, tug towards biological systems, but in the end, these intuitions might well require quite radical revision.

Intra-modular HORs

The second component of the too easy objection involves the possibility of intra-modular HORs. Rey asks us to consider a language module that "includes pragmatic aspects of utterance and comprehension" (2008, 13). Commenting on Sperber and Wilson (2002), Rey continues:

> The "meta-psychological principles" involve "presumption of relevance" that takes into account various intentions, desires and foci of attention of both the speaker and hearer, which are therefore often represented by both of them (e.g., the hearer assumes the speaker is talking about something that is *of interest* to her). That is, the module would appear to be trafficking in higher-order thoughts, but ones that speakers aren't readily able to introspect, and so [are] unconscious, as seems manifestly the case with young children (14)

In this case, it is more promising for EHOR theorists to challenge directly the appropriateness of the HORs to which Rey appeals. Such HORs would clearly not target perceptual states, they would target other *thoughts/attitudes.* But other thoughts/attitudes are not comprised of the perceptual contents upon which phenomenal character might be presumed to be parasitic. Several theorists argue that only perceptual states can be phenomenally conscious (Nelkin, 1989; Tye, 1995; Carruthers, 2005, 2011; Prinz, 2007;

2012). Several others argue that argue that phenomenal character permeates thoughts and attitudes (Strawson, 1994; Siewert, 1998; Horgan and Tienson, 2002; Pitt, 2004). The cognitive phenomenology debate is lively and unsettled, but if one takes the view that only perceptual states can be conscious (and provided one has an independent explanation), then one has a principled way of explaining why intra-modular HORs are not conscious, or at least, one has a principled way of explaining why the kind of intramodular HORs that Rey has in mind are not conscious. Ultimately, though, the question whether only perceptual states can be conscious or not is an empirical one, and can't be answered from the armchair. Thus, while the objection from intra-modular HORs puts pressure on the HOR theorist to flesh out the details of what an "appropriate" HOR is, the objection is not decisive against EHOR theories.

HORs in the Freudian Unconscious

The final component of Rey's too easy objection is the objection from Freudian psychology. The problem is this. Suppose that Freudian psychology is true. That is, suppose humans have an unconscious mind in something like the sense described by Freud, and that we repress, or the block, certain mental states from conscious awareness. Non-conscious states of guilt, say, seem to require HORs, but such HORs are non-conscious. For example,

> At a certain moment the child comes to understand that an attempt to remove his father as a rival would be punished by him with castration. So, from fear of castration...he gives up his wish to possess his mother and get rid of his father. Insofar as this wish remains in the unconscious it forms the basis of the sense of guilt. –Freud (1928/53:p229)

And similarly Rey argues that Freud:

> claimed that paranoid delusions of persecution were the result of "a person defending himself against a homosexual impulse which has become too powerful" (1917/66:p424), or, more mundanely, that jealousy is (often) the result of a man who seeks "absolution by his conscience... when he projects his own impulses to infidelity on to the partner to whom he owes faith" (1922/59:p233). As I hope the quotations make clear, all such cases would seem to involve HOTs whose targets are patently unconscious, along the lines of common self-deception, whereby people suppress unpleasant thoughts, e.g. about manifest alcoholism, spousal infidelity, or their fading youth (2008, 11-12).

Why are these cases supposed to present a counterexample to HOR theories? They are supposed to do so because, to feel guilty that one wishes to possess one's mother, say, is to have a higher-order thought (a guilt-thought) targeting one's wish to possess one's mother. But, by hypothesis of the Freudian unconscious, such a state is not conscious, even though it is the object of a higher-order thought. In fact, one doesn't even have to go Freudian here.[13] For the existence of nonconscious thoughts is now widely acknowledged by cognitive scientists, and most who do acknowledge a cognitive unconscious think that various attitudes, emotions, and perceptions can all take nonconscious form. This can all be acknowledged without invoking Freud. As Rey mentions, it is not difficult to envision a computational model of the relevant Freudian processes. For example, consider the possibility of a sophisticated Freudian machine, complete with sensory modalities that resemble a prototypical sensory modality to a sufficient degree. I already conceded that it would be plausible to attribute consciousness to a machine in possession of the latter, given that it had the appropriate cognitive architecture. But now, how would we know that such a system has *the wrong* cognitive architecture, just because it realized a Freudian psychology? In the absence of further details about the requisite kind of HOR,

[13] Rey acknowledges this point (2008, 12-13), but I think it deserves greater emphasis.

to insist that the relevant states of the Freudian machine would not be conscious begs the question.

Here there are two main methods of reply to the objection from Freudian psychology. First, a HOR theorist might tell us more about the distinctive features of HORs that are required to render a state conscious and show that HORs in the Freudian cases lack such features. Second, a HOR theorist might acknowledge inaccessible phenomenal states in select cases, providing some convincing explanation for why we should acknowledge them in those cases. Rosenthal (2005) develops the first method of response. Carruthers (2000) develops the second. I will touch on each of these in turn.

Rosenthal (2005) claims that,

> the mental attitude of a HOT must be assertoric; wondering and doubting about things, for example, do not by themselves make one conscious of those things. That's why unconsciously feeling guilty about wanting something would not alone result in that desire's being conscious, since feeling guilty about something does not involve an assertoric attitude" (185).

Commenting on Rosenthal, Rey argues that while one can surely "have a desire not to have a wish without the assertoric thought that one has it [the wish]," nevertheless, "these are plainly not the kinds of cases of (sic) Freud has in mind," because, as Freud claims the wishes still exist (2008, 12). For example, Freud says that insofar as the child's wish "remains in the unconscious it forms the basis of the sense of guilt" (1928/53, 229), and that the homophobe is defending himself against "a homosexual impulse which has become too powerful" (1917/66, 424).

It is fairly evident that the states involved are assertoric, so I won't pursue Rosenthal's method any further. Also, note that Rey's challenge here is not that the above Freudian description is in fact a true description of human psychology, but rather, that it

is "continuous with much of our ordinary understanding of people, some instances of which seem immensely plausible, and, if true, illuminating" (2008, 12). It could indeed turn out that those ordinary understandings are mistaken, however, Rey's point is that they are genuine possibilities, which HOT theory, a naturalistic theory, seems to rule out *a priori.* And again, even if you think that Freudian psychology is largely false as a description of human psychology, consider a Freudian machine. HOT theories are committed to arguing that it is metaphysically impossible for such a machine to be conscious.

To say that HOT theories rule out the possibility of a Freudian machine is somewhat misleading. The HOT theory is not intended as an *analysis* of consciousness. Rather, the constraints of the view *predict* that such a system would fail to be conscious. In the end, who knows? But more to the point, even if the HOT's in question are appropriately assertoric, as Rosenthal denies they are, their targets are inappropriate, *if* one assumes that only perceptual states can be conscious. As mentioned above, there are those who argue for such a case (Nelkin, 1989; Tye, 1995; Carruthers, 2005, 2011; Prinz, 2007; 2012). Also, as I stated in at the outset, in this dissertation I am only dealing with sensory states. I am not dealing with attitudes.

Carruthers exemplifies the second method of dealing with Freudian too-easy cases. He argues that, even if Freudian psychology were true, the Freudian unconscious is supposed to constitute an entirely distinct subject. Each person is comprised of two main subjects and so, when considering the distinction between access and phenomenal consciousness and that there are in fact (at least) these two subjects in the individual, it is

not surprising or problematic to find out that the phenomenal states of one subject could be inaccessible to the other (2000, 266-267).

On Carruthers' view, if we take seriously the idea that the Freudian unconscious (or the unconscious in our Freudian machine) constitutes a distinct subject, one that "has its own goals, beliefs, and limited powers of agency" (which is one of the premises of the Freudian argument), then it is not *quite* as surprising to think that one subject might be phenomenally conscious while, to the other, such phenomenality would be inaccessible.[14] That might seem counterintuitive at first glance, but so too does the notion of genuinely distinct subjects posited as occurring in the Freudian machine (or a human!). Once we acknowledge the possibility of genuinely distinct subjects within one human, it is not so counterintuitive that phenomenal consciousness could be severed between the two.

Thus, while the Freudian cases do present a genuine puzzle for HOR theorists, nothing can be settled in the absence of further empirical investigation, and there are things that can be said in favor of a higher-order view.

To reiterate, this section was not intended as establishing the truth of HOR theory, or as defending it from the ground up, but only to say enough to motivate the main goal of the dissertation, which is to examine differences between HOR approaches and to introduce and defend a novel version.

4. The Challenge of Higher-Order Misrepresentation

In this section I characterize the problem of higher-order misrepresentation in its standard form, as originally presented by Byrne (1997) and Neander (1998). I then

[14] For a somewhat related case consider split-brain subjects.

discuss Block's (2011) restatement of the problem. When framing the problem, it is important to keep in mind the distinction between existence versions of EHOR theory and inexistence versions of EHOR theory. The standard problem is implicitly framed as a problem for EHOR theory overall, but as we shall see, the problem does not arise in quite the same way for inexistence versions. Nevertheless, I argue that higher-order misrepresentation presents both versions of EHOR theory with a fatal dilemma. The EHOR theorist must embrace the inexistents view or what I will call a "conjunction" view. Neither of these, however, can capture the fineness-of-grain of conscious experience.

4.1 The General Problem

The problem of higher-order misrepresentation is thought to arise from the division of phenomenal labor introduced above. According to the division, a state's qualitative character is determined (and can be explained) by first-order representational properties (including causal/functional role). A state's phenomenal character is determined (and can be explained) by higher-order representational properties. EHOR theories are supposed to be naturalistic theories of mental state consciousness. But, it is plausible to think that a natural system can malfunction. Since, on a naturalistic theory of consciousness, the representational mechanisms that generate conscious states are natural ones, it is plausible to think that they can malfunction. Thus, it looks like higher-order representations can misrepresent their targets. They can do so in two main ways, each of which pose a problem for the view.

The first complication is that a higher-order state can misrepresent an *actual* lower-order target. For example, a higher-order state might misrepresent a first-order perceptual state as representing a square when, in fact, the first-order state represents a circle, *i.e.,* there might be a *mismatch* between two existing states. The second complication is that a higher-order state might represent the subject as being in a given first-order state when the subject is, in fact, *not* in any such first-order state. That is, it seems possible for a subject to undergo a *targetless* higher-order representation. For example, a subject might undergo a higher-order representation that she is in a first-order state representing a square when she is not in any corresponding first-order state at all. [15]

These two possible ways of higher-order misrepresenting are supposed to be complications for EHOR theory because they are thought to land the EHOR theorist in a dilemma. For both mismatch cases and targetless HOR cases, the EHOR theorist has the same two unsatisfying options: the EHOR theorist can either (1) deny that it would be like anything in misrepresentation cases, or (2) explain what it would be like in such cases. Option one is typically thought to be ad hoc and option two is thought to illuminate an internal inconsistency in the view. I will discuss each kind of misrepresentation case in turn.

4.2 Mismatch Cases

Consider mismatch cases first. Suppose a subject S undergoes an actual first-order state M, which represents a circle and S also undergoes a roughly simultaneous higher-

[15] Some recent higher-order theorists have appealed to data that are supposed to show that there are in fact subjects who undergo conscious experiences in virtue of targetless HOTs. Brown (2011), Brown and Lau (2011), and Lau and Rosenthal (2011). The empirical case for consciousness in virtue of targetless HOTs will be discussed in section 4.5.

order state M*, which represents M, but it misrepresents M as representing a square. The problem with denying what-it's-likeness in such a case is this. By hypothesis, the phenomenal character of M is determined by M*. Since S actually undergoes M*, we should expect M to have phenomenal character, *i.e.,* to be like something for S (being like something is what phenomenal character consists in). To deny that S would consciously experience something only in cases of misrepresentation would be ad hoc. Thus, the EHOR theorist must choose option (2) and explain what it would be like to undergo a mismatch case.

The problem with option two is that it is unclear what to say about the phenomenal character of M in such a case. According to extrinsic views, for my circle-representing state to be phenomenally conscious (have circle what-it's-likeness) *just is* for it to be appropriately represented by a higher-order state. In the case under consideration, though, that higher-order state is misrepresenting it as a square. But why then, would that make my *circle*-representing state phenomenally conscious? There is no circle (state)-representing higher-order state. The higher-order state only represents the first-order state as representing a square. Given the constraints of the view (phenomenal consciousness as a species of state consciousness, transitivity, the division of phenomenal labor), it is hard to see how an EHOR theorist could also maintain that I am aware of my circle-representing state (my circle-representing state is phenomenally conscious, or has circle what it's likeness) in virtue of my being aware of an altogether different state: a square-representing state. Here is another way to put it that brings out the peculiarity (if not explicit incompatibility) of the consequence illustrated by the previous sentence. The

extrinsic theorist must assert that my circle-representing state is phenomenally conscious (read "like circle" for me) but like square for me.

4.3 Targetless HOR cases

In a targetless HOR case, the subject undergoes a HOR with no corresponding first-order state. For example, suppose a subject S undergoes a higher-order state M*. M* represents a first-order state M. In this case, M is represented as visually representing a square, but S is actually not undergoing M. M doesn't exist, so M doesn't represent anything.

As it is in mismatch cases, option (1) is typically thought to be ad hoc as a way of dealing with targetless HOR cases. By hypothesis of EHOR theory, M* determines whatever phenomenal character S experiences. Since S really is undergoing M*, we should expect S to undergo some kind of phenomenal experience. To deny that she would only in this kind of case is ad hoc. Here again, the EHOR theorist is typically thought to be forced into choosing (2), explaining what it would be like to undergo a targetless HOR.

The trouble with (2) as a way of handling targetless HOR cases is that EHOR theories are theories of *state* consciousness, but the states that HORs render conscious are distinct extrinsic lower-order states. In a targetless case, there is no actual distinct extrinsic lower-order state to be *the* conscious state. So one might plausibly wonder which state is *the* conscious state (Kreigel, 2009). In such a case, if it must be like something for the subject, and it being like something amounts to the subject being in a conscious state, then the conscious state would have to be the higher-order state (for that is the only other relevant state the subject is in). However, this is not a conclusion the extrinsic theorist can accept;

it amounts to admitting that consciousness can be explained in terms of a single state, and that undermines the core relational strategy of extrinsic views (Byrne, 1997).

One further possibility that was mentioned in section 1 is to construe EHOR as a theory of creature consciousness, as opposed to a theory of state consciousness. On this construal a creature is conscious in virtue of what it is conscious of, but since the view is higher-order what the creature is conscious of must be a first-order state. For example, just as a pink elephant need not exist for a creature to be conscious of one (as in the case of hallucinating a pink elephant), an actual first-order representation that one is seeing the blue sky need not exist for a creature to be conscious of the blue sky. Brown (2011) suggests something like this, but he doesn't develop it much, so it is difficult to say exactly what he has in mind. One could envision what such a view would look like though. I am actually quite sympathetic to re-construing phenomenal consciousness in this way, but that is not my project here, so I will not pursue the thought any further.

Block (2011) has recently restated the problem of targetless HORs, arguing that such cases illustrate an inconsistency between the necessary and sufficient conditions of higher-order theories. Characterizing Rosenthal's version of EHOT theory, Block says "a mental state is conscious if and only if the state is the object of a certain kind of representation arrived at noninferentially" (Block, 2011: 421). This, he says, supplies EHOT with a necessary and sufficient condition for a conscious episode. The HOT supplies the sufficient condition. That there is a first-order state that is the object of a HOT is the necessary condition. The reason Block argues that these two conditions are incompatible is that in cases of targetless HOTs there is no first-order state to be the

object of the HOT. But then the necessary condition is violated, and consequently, Block concludes that HOR theories are "defunct."

4.4 Refining the Problem of Higher-Order Misrepresentation

The standard presentation of the problem, including both kinds of higher-order misrepresentation, is presumed to present a dilemma for EHOR theorists. As discussed in section 4.3, the problem is supposed to be that if EHOR theorists embrace the first horn, their answer is unavoidably ad hoc. Therefore, they must embrace the second horn. However, embracing the second horn leads to an internal tension within the view. In this section I want to clarify a few points about the standard presentation of the problem.

First, in embracing the second horn of the dilemma, the EHOR theorist is being asked to *explain* what mismatch and targetless HOR cases would be like for their subjects to undergo. While it's true that the theory should do something to help us understand such cases, ultimately, it's an empirical question and it's hard to see how any theorist will be able to explain adequately what mismatch cases *are like for someone else*, or what they *would be like for themselves* simply from the armchair, having undergone neither kind of case and in the absence of any known actual cases (a point well-stated by Lycan, 1996, 21). If they are even possible, who knows what mismatch cases would be like to undergo? Would the mismatch case described above be like circle or like square (or squarishly circle-like or circlishly square-like)? Such cases certainly seem conceptually odd, but not enough so that they can be ruled out a priori, at least not to my mind. The problem is that we currently lack the ability to determine whether or not anyone actually does undergo such states. We simply don't know if there are genuine cases.

Second, the standard presentation of the problem assumes that the first order state M must be an *actual existing state*. That is, the targetless HORs problem is supposed to be a problem because there is no state to be *the* conscious state. But stating the problem that way assumes that M must be actual, and that assumption only poses an immediate problem for *existence* versions of EHOR. Recall, though, that Rosenthal and Weisberg defend what I called the "intentional inexistents" version of EHOR theory. On their view, a conscious state may be underwritten by an intentionally *in*existent first-order state (M doesn't have to be an actual state the subject undergoes). Since Rosenthal and Weisberg explicitly reject the existence condition, they need not commit themselves, for example, to Block's reading of the biconditional as an accurate characterization of their version of EHOR theory. And if it is not necessary to have an actual lower-order state in the first place, then the targetless HORs objection doesn't get off the ground against their view, or at least not in the same way (below I will argue that a different problem arises).

Finally, as a matter of fact, actual EHOR theorists do happen to choose the second horn of the dilemma, but it is not clear that denying what-it's-likeness in higher-order misrepresentation cases is inescapably ad hoc, so it isn't clear that they must choose the second horn. For example, one might develop a "conjunction," or "joint determination" EHOR theory.[16]

One way to develop such a view more clearly is to set matching higher-order and first-order components as a constraint for phenomenal consciousness. Without further

[16] Lau (Lau and Brown, forthcoming) is one of the few theorists to hint at a joint determination view, but what he means by 'joint determination' is not entirely clear. As he describes it, his view could be taken as a mere conjunction view or a kind of intrinsic higher-order view.

explanation, this would be ad hoc. However, there is further explanation that can be given. The traditional extrinsic HOR theorist already endorses: the claim that phenomenal consciousness is a species of mental state consciousness, the transitivity thesis, and the division of phenomenal labor. But then the denial of mismatch cases seems virtually entailed by these constraints. For example, while the division of phenomenal labor divvies up accounts of qualitative and phenomenal character, one might plausibly argue that phenomenal character is, in some way, parasitic upon qualitative character. That is, it is not that first-order representational properties determine qualitative character and higher-order properties determine phenomenal character *and* qualitative character (in the absence of an actual first-order state, say), but rather, that first-order representational properties determine the qualitative character (the qualitative features represented in the conscious state, *e.g.,* blueness) and higher-order representational properties determine *only* the phenomenal character (what that blueness is like for the subject). But since phenomenal consciousness depends in part on there being qualitative character, if there is no first-order state, then there is no qualitative character to partly constitute the phenomenal character. Thus, if there is no *matching* first-order state, then it won't be like anything for the subject. The *relevant* qualitative component will be missing. For in addition to the seeming aspect of phenomenal character, one also needs the qualitative aspect to seem like something. In targetless cases the qualitative component is completely absent. In mismatch cases the qualitative component is effectively absent. This method of reply provides a way for the EHOR theorist to handle both mismatch cases and targetless HOR cases, and the method is not ad hoc. Rather, the requirement that the first and

higher-order components match is a consequence of the more fundamental constraints of EHOR theory. It actually has theoretical force.

Thus, it looks like there are two more refined options for responding to the problem of higher-order misrepresentation. EHOR theorists can endorse a conjunction view, embracing the first possible path of explanation, or they can endorse the intentional inexistents view, embracing the second possible path of explanation. Both of the options have some initial attractiveness, but I will argue that, upon further reflection, they both fail for similar reasons. For neither the conjunction view nor the inexistents view can accommodate the fineness-of-grain of conscious experience.

4.5 The Argument from Fineness of Grain

Individual percepts are commonly described as having a determinacy of detail that exceeds one's conceptual repertoire. This determinacy of detail is fineness-of-grain (or FoG, for short). For example, the specificity of the various shades of green my experience represents in the trees outside my window or the specificity of the various shapes of leaves, branches, or objects on my desk far exceeds any description that I might attempt to give. For these things are kind of roundish/squarish, &c., but those descriptions fail to capture the specificity of detail represented in my experience. Several of those shades and shapes are properties for which I have no concept.[17]

[17] Contrast FoG with what is often described as "richness." Whereas FoG applies to individual percepts (the specific shade of yellow of a tennis ball), richness is presumed to be a characteristic of whole experiences (the multitude of details in the vista before me). For example, as I stare out at the vista I see many things. Those who think that experience is rich say that I experience phenomenally more than I am able to access, or notice at a given conscious moment (Tye, 2006; Block, 2001). Those who deny richness claim that we are blind to the features that others allege to be phenomenally conscious

In this subsection I will argue that fine-ness of grain presents the EHOR theorist with a dilemma. The reason FoG poses a problem for the conjunction view is this. On the conjunction view, phenomenal consciousness is jointly determined by a higher-order state and its extrinsic first-order target. One motivation for the view is the notion that phenomenal character is parasitic, in some way, upon qualitative character. If there is no first-order qualitative state, then there won't be phenomenal consciousness. If there is a mismatch, the first-order state is effectively missing. This might offer a way of responding to higher-order misrepresentation complications, but it re-saddles the view with the challenge of the other other-minds problem for cases of *veridical* higher-order representation. This reason why it does do is that an appropriate HOR in one person can target an appropriate FOR in another person. On the face of it this seems absurd, however, the conjunction view doesn't tell us why it is; it doesn't tell us why we should consider only intrasubjective M*/M pairs "conjunctions," but not intersubjective M*/M pairs.

Moreover, intrasubjective conjunction itself seems insufficient for another reason. Recall that on the conjunction view M* and M are numerically distinct states. One might wonder why a distinct extrinsic representation that targets a first-order state would make it like anything for its subject in the fine-grained way that conscious experience seems to be. Again, if an appropriate HOR in person A targeting an appropriate FOR in person B doesn't render person B's state conscious, why would an analogous pair do so intrasubjectively? One possible answer is that there is a tighter connection between the two components in intrasubjective cases, which is absent in other-minds cases.

---

but unaccessed (O'Regan and Noë, 2001). While the richness issue has been widely debated, I know of no one who denies that experience is finely grained in some sense.

One might try to cash out the "tighter connection" in terms of an EHOP theory. On such a view, the inner sense organ of one person does not have access to the outputs of another person's first-order sensory systems. Thus, the rejection of other-minds cases might well be fairly straightforward. I am not going to discuss this possibility, though, because presently there is no reason (at all) to think that there is in fact an inner sense organ, which scans sensory systems. Maybe there will turn out to be such an organ, but right now I'm not going to pursue the consequences of that possibility.

On the EHOT theory, the higher-order state is an assertoric thought. It is hard to see why an assertoric thought *that* a subject S is undergoing an appropriate first-order state will make it like something for S, even if the HOT arises non-inferentially (or at least, is generated by processes which are not consciously inferential). Having an assertoric thought *that* one is in M might make S aware *that* she is in M (which has such and such content), but one may still wonder whether it would make S aware of M's content, such that it is experienced by S, rather than merely acknowledged, or known of. The fineness of grain that seems to partly constitute conscious experience is part of a distinct extrinsic state.[18] The conjunction view needs fineness of grain to deny what-its-likeness in misrepresentation cases (it's the FoG that's missing in such cases), but FoG is exactly what comes back to haunt the view in veridical cases. For even in veridical cases the FoG is, in an important sense, inaccessible.

[18] This distinction (and the task of making sense out of it) goes back at least to Russell's distinction between "knowledge by description" and "knowledge by acquaintance." To put the distinction a bit differently, one might think that an assertoric thought about an appropriate FOR would make the subject aware that she is undergoing it in the descriptive sense, but not in the acquaintance sense, even though the first-order state is fine-grained.

The reason FoG poses a problem for the inexistents view is that experience seems finely grained. This feature needs explaining, even if it is just the *appearance* of fineness of grain that gets explained. In fact, acknowledging the FoG of conscious experience is exactly what enabled the conjunction theorist to deny phenomenal consciousness in targetless HOR cases. For if there is no finely grained first-order representation, but merely a higher-order thought *that* there is such a state, then there won't be an appropriate FOR to be conscious of. The inexistence theorist, on the other hand, claims that there is phenomenal consciousness even in targetless HOR cases. As such, the inexistence theorist must either explain how there could be FoG in targetless HOR cases or deny that experience is finely grained.

It is difficult to see how the inexistence theorist will be able to account for FoG, though, for similar reasons that complicated the conjunction view. The problem with explaining FoG is that on the Rosenthal and Weisberg inexistents view, the requisite kind of higher-order state to issue in phenomenal consciousness is an assertoric thought. An assertoric thought, though, is coarsely grained, at least more so than perceptual experience. A mere assertoric higher-order thought on its own will not even be partly constituted by perceptual contents (even its first-order component), and it is from perceptual representation that FoG might be plausibly thought to derive. If you take FoG seriously, and it is at least plausible to think this feature of mental appearance ought to be respected, then the inexistence theorist must either deny FoG or deny phenomenal consciousness in targetless HOR cases. Denying phenomenal consciousness in targetless HOR cases is not an option for the inexistents view (that there is phenomenal consciousness in targetless HOR cases is the core component of the view). Thus, it looks

like the inexistents theorist must deny FoG altogether. Denying FoG is certainly an option, but it seems ill-motivated, since one aim of the view is to explain mental appearance. I think everyone agrees that conscious experience at least *appears* finely grained, even if that appearance is an illusion.

The problem that emerges in the veridical cases is this. Recall that Rosenthal and Weisberg agree that the two kinds of case would be phenomenally the same. They think that a targetless HOT will be just like an identical HOT, which targets an actual first-order state. In veridical cases of higher-order representation, the HOT corresponds to an actual first-order state, but the HOT is still just an assertoric thought about an extrinsic state. Moreover, even if the intentional object happens to correspond to an actual first-order state, the conscious state is still supposed to be an intentional object. The conscious state is an "intentional existent." If so, one might wonder what difference undergoing the actual first-order state could possibly make and why it would make any difference. That is, on the conjunction view, the problem was that the mere conjunction of two actual states is insufficient. But the inexistents view asserts something rather more puzzling: it is not the conjunction of two existing states that really matters, it is the conjunction of an actual higher-order state and a purely intentional first-order state, whether that first-order state really exists or not. Given the "distance" between a higher-order state and its actual object, I argued that the conjunction view is problematic, on the grounds that the FoG that is supplied by the first-order state is not "undergone." Rather, it is something that is known of (or known about) via a distinct extrinsic HOT. The "distance" between the FoG-bearing state and the HOR is even greater for the intentional inexistents view. In that case, the FoG isn't even part of the putative conscious state itself (the intentional

object), rather it's a constituent of an *actual* first-order state, which even the inexistents theorists acknowledges the subject is not undergoing.

Rosenthal has argued that FoG can be accounted for solely in terms of *the way* a particular HOT represents a first-order state to its subject, which, for him, amounts to the concepts that are deployed as part of the HOT. For example, when I undergo an appropriately assertoric HOT asserting that I am experiencing a certain shade of red, with a certain saturation and brightness, that thought itself determines what it's like for me to undergo the experience described (2005, 186). Rosenthal thinks that FoG is exhausted by such thoughts.

I agree that concepts typically contribute to the qualitative and phenomenal character of conscious experience. But the problem with this aspect of Rosenthal's view is that the HOTs to which he appeals are merely descriptions; they are merely concepts. As argued above against the conjunction view, it is hard to see why an assertoric thought *that* a subject S is undergoing an appropriate first-order state will make it like something for S, even if the HOT arises non-inferentially (or at least is generated by processes which are not consciously inferential). Again, having an assertoric thought *that* one is in M might make S aware *that* she is in M (which has such and such content), but one may still wonder whether it would make S aware of M's content, such that it is experienced by S, rather than merely acknowledged, or known of. The fineness of grain that seems to partly constitute conscious experience is part of a distinct extrinsic state.

Rosenthal also claims that experience becomes more finely grained in virtue of "verbal pegs on which to hang those conscious experiences" (187). Concepts like TANNIN, SHARP, OAKEY can all serve to enhance the experience of wine, making it

more finely-grained. Here too I agree with Rosenthal that such concepts enrich experience, but they do so only in tandem with the first-order properties to which they refer. Without those first-order contents the concepts are, to borrow Rosenthal's analogy, just empty pegs (compare the concepts in isolation to the situation of Mary, before she leaves her black and white room. She has a concept of red, but it lacks the appropriate first-order content that can be acquired only through undergoing the experience itself, or at least that is how I will characterize the situation in Chapter 4).

Finally, suppose one endorses a view like McDowell's (1995), according to which fine-grainedness is characterized in terms of indexical concepts. As applied to Rosenthal's account, on such a view there would seem to be nothing excluded by the conceptual content of a HOT, where these bottom out in indexical concepts. It won't help Rosenthal to characterize the content of the HOTs in terms of indexicals that point to properties of first-order representations, though. For Rosenthal the HOT is supposed to determine phenomenal character on its own, even when no actual FOR is present, but the idea that indexicals pointing to nothing could constitute the phenomenal character of even a simple non-comparative experience of green is implausible. An indexical that points to nothing has no character at all.

The point that I have been driving home is that we seem to need actual first-order representations in addition to higher-order representations of them. But again, the FORs need not be veridical. One needs to undergo an actual appropriate FOR, an actual experience (broadly construed to include hallucinations, illusions, and ordinary perceptual errors), to be thought about. This pushes us back towards a conjunction view,

but as I have argued above, the mere conjunction of the two states is insufficient. Thus, it looks like the two best options for the EHOR theorist are both inadequate.

4.6 An Empirical Case for (Some Kind) of Higher-Order Theory?

Some theorists have recently appealed to various empirical data to try to firm up the case for consciousness in the absence of actual first-order states (Lau and Rosenthal, 2011; Lau and Brown, 2011). However, it is extremely questionable that the data they cite supports consciousness in the absence of any actual first-order states at all.

One intriguing case involves subjects with Charles Bonnet syndrome (CBS). Subjects with one form of CBS have damage to some portions of the early visual processing regions of the brain, *e.g.,* V1 (Ashwin and Tsaloumas, 2007). On one view (the feedback-to-V1 view), V1 houses the neural correlate of first-order representations, which is exactly the area damaged in CBS cases. Nevertheless, CBS subjects report robust conscious visual experiences, including vivid hallucinations of faces, familiar persons or objects, and complex geometric patterns. And, importantly, CBS subjects are thought to be cognitively intact: the subjects have no known additional damage and they can lucidly describe their hallucinations while accepting that the visions are the result of visual deficit. Thus, if one were to endorse the feedback view, one must claim that in CBS cases there are no "first-order" visual representations at work.

From the perspective of higher-order theorists, CBS cases would seem to present a puzzling result, but one that is no doubt desirable to the inexistents view. By hypothesis of EHOR theory, for a state to be conscious is for it to be appropriately represented. In such cases, what would seem to be the relevant first-order state is missing, but yet

subjects still report phenomenal experience. One might conclude, then, that CBS cases provide actual examples of consciousness in virtue of a targetless HORs. The conscious state, one might conclude, is an intentional inexistent.

The main problem is that the argument relies on a contentious assumption about the neural correlate of first-order representation and from there, assumes that the phenomenal work is being done by a higher-order state. The area assumed to be the correlate of first-order representation is the feedback circuit from extrastriate regions to primary visual cortex (V1).[19] For the main cases cited are cases wherein the subject has undergone damage to V1. The problem is that, while the feedback view has its proponents (*e.g.,* Lamme, 2006; and *possibly* Block, 2005, 2007), it also has its opponents (*e.g.,* Macknik and Martinez, 2007; Silvanto and Rees, 2011; Prinz, 2012). The truth of the empirical case for consciousness in virtue of targetless HORs depends, to a large extent, on whether the specified region is, or is not, the neural correlate of "first-order representation," something that is presently undetermined and relies on further investigation.

Moreover, there is a very plausible case to be made in favor of the contrasting view that first-order representations should be correlated with extrastriate regions, rather than with feedback loops to V1. On that view, CBS cases present less of a puzzle, because cortical areas outside of V1 are still active. The upshot of the targetless HORs data is that

[19] While Lau and Rosenthal seem to operate on this assumption. It is not entirely clear what Lau and Brown think. At one point they say they "consider feedback to V1 as the primary candidate for the correlate of first-order representations" (2011, 5). In a later passage in the same work, they say they "think locating first-order representations in extrastriate cortex is superior to the feedback-to-V1 view" (12).

the current data do not decide much of anything. In particular, there is still no solid empirical case for consciousness in virtue of targetless HORs.[20]

5. Conclusion

Higher-order theories present a promising naturalistic strategy for explaining some of the distinctive features of phenomenally conscious states, but they have faced a plethora of objections. Some of these are objections to which traditional extrinsic higher-order views can adequately reply. There is one, however, that I have argued is an outstanding problem for the view. This is the problem of higher-order misrepresentation. For those who remain sympathetic to the higher order framework, the problem of higher-order misrepresentation in particular has motivated a modified higher-order approach. This is the self-representational theory of consciousness, or what I call *intrinsic higher-order theory* and it will be the focus of the remaining chapters.

[20] There are other cases considered by Lau and Brown (forthcoming) which aim to show that there can be phenomenal consciousness when there are in fact first-order states, but when these states are judged to be "too weak" to underwrite the robust phenomenology reported by their subjects. These cases are interesting, however, they do not bear directly on the phenomenon of higher-order misrepresentation and, for that reason, I will not consider them here.

Chapter 3: Intrinsic Higher-Order Theory and Higher-Order Misrepresentation

Despite the many virtues of the higher-order framework, existing extrinsic theories face substantial objections. Proponents of the four leading intrinsic higher-order views are motivated in large part by these objections, especially higher-order misrepresentation.[21] This chapter examines the four leading versions of intrinsic higher-order theory. These are Rocco Gennaro's "wide intrinsicality view" (1996, 2004, 2006, 2012), Peter Carruthers' "dual-content theory" (2000, 2005), Robert Van Gulick's "HOGS model" (2004, 2006), and Uriah Kriegel's "cross-order information integration hypothesis," (2005, 2009). The central claim of the chapter is that the leading versions encounter difficulties addressing higher-order misrepresentation. They either fail to address it altogether or inherit further complications in attempting to address it. Weisberg (2008) has argued similarly. In contrast to Weisberg, I argue that an appropriately modified intrinsic HOR theory can address the problem of higher-order misrepresentation. In Chapter 4 I begin to develop that modified view as a version of intrinsic higher-order thought theory.

1. Existing IHOR Theories

Recall that according to the extrinsic higher-order theorist, a mental state M of a subject S is conscious if and only if M represents and is itself represented by a

[21] See, for example, Gennaro (1996, 2012), Carruthers (2000), Van Gulick (2006), Kriegel (2009)

numerically distinct, extrinsic state M* (even if, on some views, M is an intentional inexistent). The core assumption of self-representational, or what I am calling "intrinsic higher-order" theory, is that a conscious mental state is a complex mental state representing both the world and itself (or at least one of its own parts).[22] More precisely, a conscious mental state has two critical components: a lower-order component, representing some feature of the subject's environment (construed broadly to include the subject's body) and a higher-order component, representing the lower-order component. Thus characterized, one can see that self-representationalists build upon the foundation laid by traditional higher-order theorists.

The ultimate explanatory success of IHOR depends on the details of the specific version in question. These details will be the focus of the section. First, notice that the general intrinsic framework inherits the explanatory power of extrinsic higher-order theory. For example all intrinsic theorists agree that conscious states require appropriate HORs and appropriate FORs. This alone preserves the explanatory power of extrinsic views regarding D1-D4 from Chapter 1. It seems, then, that there is nothing lost in tightening the relationship between the first-and higher-order components. Indeed, intrinsic theorists argue that tightening the relationship increases the explanatory power of the EHOR framework. Intrinsic views, they argue, can account for intimacy and explain away higher-order misrepresentation, two outstanding worries for EHOR views.

[22] There are various ways in which the relation between the relevant parts of a conscious state might obtain. See Kriegel (2006) for a survey of the possibilities.

1.1 The Wide Intrinsicality View

Gennaro's "wide intrinsicality view" (WIV), is a version of intrinsic higher-order *thought* theory. Here is how Gennaro describes the view:

> My WIV does not treat the conscious rendering state as entirely distinct from CS [the conscious state]… Rather, it treats conscious states as complex states with both CS and the meta-psychological states as parts. Conscious states are individuated widely so as to treat the meta-psychological state as intrinsic to the conscious mental state' (1996, 16).

That is, contra extrinsic higher-order theorists, who contend that the conscious state of an M/M* pair is the lower-order state M, Gennaro claims that we should treat the two states as parts of a single complex state with both first-order and higher-order (or "meta-psychological") components. Continuing, Gennaro writes

> On the WIV, we have two parts of a single conscious state with one part directed at ("aware of") the other. In short, there is a complex conscious mental state with an inner, intrinsic relation between parts (2004, 60-61).

Most recently,

> According to the WIV, what makes mental states conscious is *intrinsic* to conscious states, but a kind of *inner* self-referential and relational element is also present *within* the structure of such states. In contrast to standard HOT theory, the WIV says that *first-order* conscious mental states are *complex* states containing both a world-directed mental state-part M and an unconscious metapsychological thought (MET). It is, if you will, an intrinsic version of HOT theory (2012, 55).

Thus, on Gennaro's WIV, what makes a conscious state conscious is that one part of the state (the higher-order part) is directed at another part (the lower order-part).

Kriegel (2005, 2006) maintains that Gennaro's theory makes consciousness purely notional, since according to Gennaro we are merely to "treat," both states in conjunction to be the conscious state. If that is the case, then Gennaro's view smacks of arbitrariness; we could "treat" almost any two states we so desire to be a single state of

the system. However, while it is true that if we were merely to "treat" the two states as one, the view would smack of arbitrariness, construing the WIV this way is uncharitable. It is clear from the above passages (and others) that Gennaro at least intends the higher-order and lower-order components to be "importantly related" (1996, 16), not just notionally related. It is also clear that the conscious state is intended to be a complex rather than a conjunction. Whether or not Gennaro can make good on that claim is a separate issue. Genarro's WIV is not as trivial as Kriegel presents it. The greater worry is that the account of the more robust bond between these two crucial components is wanting, and this threatens to weaken the view.[23]

To explain the relation between the lower-order and higher-order component of a conscious state Gennaro invokes a quasi-Kantian notion of "synthesis." Echoing Kant, he says, "the understanding unconsciously 'synthesizes' the raw data of experience" (2006, 237; 2012, 78). Just as concepts, according to Kant, are presupposed in experience, for Gennaro "the *concepts* that figure into the HOTs are presupposed in conscious experience," they are what make conscious experience possible (2012, 77). Gennaro's basic idea is that we first receive information via the senses. This is early perceptual processing of what Gennaro calls the "raw data of experience." Then, some of it rises to the level of nonconscious experience, and those nonconscious experiences do not become conscious until one applies concepts to them. Gennaro thinks the application of concepts should be understood in terms of HOTs being directed at such information. Continuing he writes,

[23] Notice too that the extrinsic theorist also asserts that a higher-order state and its lower-order target must be "importantly related" somehow. See, *e.g.,* Rosenthal (1997, p. 744). Part of the task of furthering the HOR view is to elaborate on the "important relation."

> I experience the brown tree *as a brown tree* partly because I apply the concepts "brown" and "tree" (in my HOTs) to the incoming information via my visual perceptual apparatus. More specifically, I have a HOT such as "I am seeing a brown tree now" (78).

Being constituents of the same state, then, is the "important relation" that Gennaro has in mind and "synthesis" is supposed to explain the process of integrating the two relevant components of a conscious state together. I take it that being synthesized is supposed by Gennaro to be a psychologically real process, which is not merely notional (something which matters to Kriegel, as will be illustrated in section 2.4).

If the WIV were successful, it is relatively clear how it would avoid the rock objection. If synthesis can be appropriately cashed-out, it certainly seems like rocks cannot be synthesized into a complex mental state. A rock can't become a proper part of a mental state. Similarly, since it doesn't look like anyone else's mental state can be synthesized into one of my complex states, the WIV might have a plausible way of avoiding the other other-minds problem.

It is also fairly clear that, if the WIV were successful, the too-easy problem would not present quite the same challenge that it does for certain versions of EHOR theory. Consider the objection from the Freudian unconscious. First, there is nothing about the Freudian unconscious that requires a "synthesized" state, at least not on Gennaro's construal of synthesis. All one needs to run the objection from Freudian psychology is to posit non-conscious "appropriate" higher-order representations, but not complex, or synthesized ones. Though, for similar reasons as those that were discussed in Chapter 2, the WIV theorist will most likely have to concede that an appropriately modified computer is at least capable of being phenomenally conscious (unless there is something specifically human, or biological, about synthesis, which, at least to my mind, seems

unlikely). Of course, we could always rerun the Freudian objection, positing the occurrence and explanatory prevalence of synthesized states instead of HOTs. Run this way, the objection falls short before it gets going. For while there will surely be "synthesis" of first-order concepts and the raw data of experience in Freudian cases, it seems plausible that the synthesis of higher-order concepts and raw data is exactly what is missing from such cases. Indeed, synthesis might actually offer one explanation for why Freudian cases fail to be conscious. The WIV may or may not have something to say about that, however, since the view suffers from a much deeper problem, I won't pursue any of its possible replies here.

Gennaro clearly thinks the view rules out higher-order misrepresentation. He writes,

> There is…a kind of infallibility between [the first-order and higher-order components] on the WIV…The impossibility of error in this case is merely within the complex CMS [conscious mental state], and not some kind of certainty that holds between one's CMS and the outer object (2006, 242-243).

On the WIV, then, you simply can't get a mismatch between the two critical components of a conscious state. It's not that you can't misperceive some external object, but rather that, once integrated, you can't higher-order misrepresent the state that has been integrated. That's just what a conscious state is, on Gennaro's view.

Notice Gennaro's leap from concepts such as "brown" and "tree" to the claim that they should be construed as *higher-order* thoughts. It's not at all obvious that the application of such concepts must be, or even is, higher-order in the relevant "metapsychological" sense that Gennaro thinks they are. As argued above against Rosenthal's EHOT view, concepts such as "brown," "tree," "oakiness," and "tannin" don't seem higher order at all. These concepts apply to the world, to the presumed

properties of the objects to which they refer (wine and trees), not to the mental states that represent them. The higher-order concepts Gennaro needs are concepts such as "seems brown," "seems tree-ish," &c. But Gennaro gives no explanation for why we should think that these are the concepts that are, in fact, at work in the theory. He merely slips from first-order talk to higher-order talk, appealing to the Kantian notion of synthesis to support the structure of conscious states that his own view posits.

Sometimes Gennaro describes the situation a bit differently. Sometimes he claims that the above mentioned concepts (brown, tree) are *contained in* HOTs, *i.e.,* the relevant concepts are not merely first-order such as BROWN, OAKEY, &c. are, but rather, they are first-order concepts integrated into higher-order ones, *e.g.,* I AM EXPERIENCING BROWN. However, even if we grant that the concepts to which Gennaro appeals are higher-order, or that they are contained in higher-order thoughts, we still have no explanation for the integration of FORs and HOTs. Why should we think that synthesis actually binds two *states* together in the way that Gennaro claims it does? No answer is provided.

One of the main worries about the WIV is that Gennaro does not provide much of an account of synthesis. All he really says is that neural feedback loops and a conscious complex might be coinstantiated. But the coinstantiation of feedback loops does not *explain* synthesis. For synthesis is supposed to be a process that produces a state with intimate relations between its parts. In particular, neural feedback loops do not tell us anything about how synthesis integrates a higher-order thought with its first-order target (the "raw data of experience").

Weisberg argues that there appear to be nonconscious cases where the very same operation of synthesis is at work. For example, perceiving money *as* money partly explains how we disambiguate the word "bank" in masked priming tasks. Since subjects report no awareness of the priming image, while behaving in ways that indicate the priming image was visually processed and categorized, this strongly suggests that some kind of "synthesis" of concepts and the "raw data of experience" are operating, but they are doing so in the absence of consciousness (Weisberg, 2008, 173). However, it is fairly straightforward to envision two kinds of synthesis, *viz.,* first-order synthesis and higher-order synthesis. So the fact that synthesis operates at the first order, when such states aren't conscious on its own is not a problem for the view.

The main problem for the WIV is that we still need an explanation for how synthesis could rule out higher-order misrepresentation. Here is why. Recall Gennaro's appeal to feedback loops and the synthesis of concepts with the raw data of experience. Feedback loops alone do not provide a reason to think that higher-order misrepresentation would be ruled out, but neither does synthesis, as described by Gennaro. For example, neither tells us why a lower-order state representing a circle and a higher-order representation of that state as representing a square could not be synthesized into the same global state. This becomes especially noticeable when one appreciates the analogy that Gennaro sets up between higher-order synthesis, and the Kantian notion, which seems to be first-order synthesis. The first-order integration of concepts and perceptual content involves categorization, but there is nothing about the integration of first-order concepts with first-order perceptual content that rules out miscategorization, or misrepresentation. Indeed, any plausible account must accommodate first-order

misrepresentation, since we do indeed undergo states wherein we misperceive. But then, why should we think that the analogous synthesis procedure, one level up, would do any better? It seems that there is still nothing to exclude the possibility of a higher-order concept miscategorizing its first-order target. And if there is not, then the view is still subject to the challenge of higher-order misrepresentation.

The fundamental problem for the WIV is that, as it stands, it looks like it is founded on a stipulation, *viz.* the stipulation that conscious mental states simply cannot involve mismatching first-order/higher-order pairs. But the only explanation we get for why this would be so, is that conscious mental states are formed via synthesis. And again, nothing about synthesis, as characterized by Gennaro, rules out mismatches. Thus, if synthesis cannot provide the requisite explanation ruling out higher-order misrepresentation, then all we're left with is a stipulation: That's just how conscious states are structured. But we don't want a definition. We want an explanation.

Overall, while the WIV gestures in the right direction, it leaves synthesis unexplained, which is the primary mechanism purported to do the critical explanatory work of the view. And even if we grant that synthesis, as described by Gennaro, is plausible, it still does not rule out higher-order misrepresentation.[24] Thus, if higher-order misrepresentation is a genuine problem for EHOR theories, Gennaro's proposed improvements do not solve the problem.

[24] In Chapter 4 I will invoke something like Genarro's notion of synthesis. However, I will not call it that and I will attempt to give the process a more mechanistic characterization

1.2 The Higher-Order Global States Model

As its name suggests, Van Gulick's "higher-order global states" (HOGS) model, has two main components: a higher-order component (HO) and a global states component (GS). The global states component derives from Baars "global workspace theory" (1988, 1997) and Dennett's notion of "cerebral celebrity" (1991, 1992, 1995a, 1995b, 2001).[25] Van Gulick maintains that a lower-level state is "recruited" into a complex global state. The first-order state then becomes conscious, partly in virtue of being integrated into that global state and acquiring new connections to other states and processes. Since the conscious state is intrinsic to the global complex, which is, for Van Gulick "higher-order" (more on exactly why it's higher-order below), the view is a kind of intrinsic higher-order theory: both the lower-order and higher-order components are parts of the same (global) state.

Van Gulick's characterization of meta-intentionality differs from standard characterizations. Typically, it is claimed that the higher-orderness of a state is a matter of the state representing another mental state, but the higher-order representation has more demanding constraints on standard views. That is, while on standard views a subject need not form a thought with the exact structure "I am experiencing *x,*" and, while the subject need not be conscious of that higher-order thought, higher-order representation is (cognitively speaking) a higher-level phenomenon, perhaps propositional in structure (or on Lycan's view requiring a higher-order perceptual

[25] See also Baars, Ramsøy, and Laureys (2003). For Baars, global broadcast was supposed to explain consciousness, and perhaps it does explain *one* kind of consciousness, *viz., access* consciousness, or that a mental state is conscious in the sense that it is accessible to inform reasoning, decision-making, and generate verbal report. However, most philosophers deny that Baars theory explains phenomenal consciousness.

component generated by inner sense). On Van Gulick's view, meta-intentionality is certainly present in conscious states, but it is also present in a wide range of what would typically, according to standard views, be considered lower-level, non-conscious states.

Van Gulick thinks that lower level states are "implicitly" meta-intentional. Sometimes Van Gulick describes such states as having implicit "reflexive self-awareness," (2006, 24). Sometimes he describes them as having implicit "self-understanding." In earlier work the idea seems inspired by Kant's notion that there is an intimate connection between consciousness and self-awareness (2004, 84). In later work, Van Gulick attempts to derive the idea from what he calls the "teleopragmatic" view of mind, which highlights the mind as a biological entity, the primary purpose of which is to enable an organism to interact successfully with its environment. To do that, the mind must be representationally, or intentionally "tuned" to the things in its environment with which it interacts.

Van Gulick then invokes notions of 'information', 'content', and 'understanding' in rather elemental forms. On the teleopragmatic view, he writes,

> [o]rganisms can be informed about or understand some feature of their world to many varying degrees. All such cases involve nonrandom correlation or tracking between the organism and the features about which it is informed, (2006, 19).

For example, a bat's wings implicitly carry information about the bat's environment, because they have been adapted, for example, specifically to the air through which the bat flies. They (the wings, not the bat!) *implicitly represent* certain features of air (18). On the teleopragmatic view a bee "understands" the correlation between a particular flower's fragrance and its nectar "insofar as its system of behavioral control guides it along the gradient of scent that maximizes its foraging success" (18). On Van Gulick's view, it

seems that any animate organism (or part of an organism!) that has evolved represents, or can represent, at least in some minimal sense of 'represent' (plants, fish tails, panda's thumbs).

Importantly, on the teleopragmatic view, not only is there a wide range of different things that represent, but also, there is a wide range of possible content, which can vary in both type and sophistication, and, just as bat's wings or bees are characterized as "understanding" certain things, so too are various lower-level sub-personal structures in minds. However, such structures are not only tuned to external things in an organism's environment, they are also tuned to the internal representational states and structures with which they interact. Insofar as these structures are tuned to internal states, they have what Van Gulick refers to as "implicit self-awareness." It is in this regard that Van Gulick thinks lower-level nonconscious mental states are meta-intentional, even though they are not conscious. The sense in which one lower-level state is about another, or functions as the content of another, or understands itself, is the teleopragmatic "tuning" sense, introduced above. Such states are meta-intentional insofar as they have "specifically adapted to the intentional nature and content of the states and processes to which they apply," but such states need not be as cognitively complex as extrinsic higher-order theorists tend to assume, *e.g.,* they are exemplified by the kinds of sub-personal states that underlie certain basic learning processes (Van Gulick, 2006, p. 21-22). Even low level states involve an element of implicit self-awareness.

Van Gulick maintains that it is only in virtue of the fact the we "embody such a rich store of implicit and procedural self-awareness at the subpersonal level," that we can be conscious (23). There's meta-intentionality all over the phylogenic scale, and,

throughout a wide range of our various cognitive states and structures, but these states/structures are not conscious. On the HOGS model, meta-intentionality itself is not sufficient for consciousness. What you need is the appropriate *degree* of meta-intentionality, which is acquired in virtue of global accessibility.

With the abovementioned view of meta-intentionality in place, Van Gulick appeals to the "global workspace" theory of consciousness. On the most developed global workspace theory, the mind is organized around the global availability of information in the brain to specialist subsystems. The result of global availability is to make the contents widely available to various processing systems, which, according to such theories, thereby makes the states representing those contents conscious.

The two main components of the HOGS model come together to explain consciousness in the following way. When a first-order state embodying a rich store of implicit self-awareness is recruited by the global workspace, it attains a higher-degree of implicit self-awareness. It does this, not in virtue of being explicitly represented by a distinct state, or even by one of its own parts, but rather, merely in virtue of having meta-intentionality built into its structure, which, once recruited, is part of a global state. In virtue of that, the first-order state is transformed to a conscious state. Van Gulick writes, "The transformation from unconscious to conscious state is not a matter of merely directing a separate and distinct meta-state onto the lower-order state but of 'recruiting' it into the globally integrated state," (2004, 74-75).

The HOGs model has a straightforward response to both the rock problem and the higher-order misrepresentation problem. On the HOGS model, the first-order state becomes conscious in virtue of being recruited in a global state, so the issue of whether a

rock would be rendered conscious is a nonstarter; rocks cannot be integrated into mental states.

Moreover, because the original state is recruited into the global state, higher-order misrepresentation seems ruled out, because that very activation and the content realized by its states are parts of the global state.[26] It would also provide a straightforward answer to the question introduced above that the extrinsic theorist seemed unable to answer: when targetless higher-order representations arise, which state, exactly, is *the* conscious state? Van Gulick's answer is that in such cases, there is no conscious state, because there is no state that gets globally integrated in the first place. If there is no first-order state to contribute *its* degree of implicit self-reflexive awareness, then there is no state to take on a heightened degree of implicit self-reflexive awareness. But notice that Van Gulick's conclusion that you cannot have a conscious state without a first-order state is just that: it is a conclusion, which follows from an *explanation*. It is not a stipulation, as I argued it is for Gennaro's WIV.

The HOGS model obviates the rock problem and rules out higher-order misrepresentation straightforwardly. The crucial question, then, is: can the HOGS model handle the too easy problem? And it looks like the model has a response to part of Rey's too-easy problem, but still suffers from one component of it.

---

[26] Van Gulick does allow that the recruited state could be altered in virtue of being integrated into the global complex, but that doesn't entail higher-order misrepresentation. For the state recruited forms a critical part of the complex global state and will determine the first-order character of the global state. There cannot be a mismatch, because the implicit higher-order content of the global state is partly constituted by whatever the recruited state represents (post-alteration).

Recall that two components of the problem were the threat of HORs in ordinary computers and intramodular HORs. The reply considered in Chapter 2 was that an adequately developed HOR theory will set constraints on the kinds of first-order representational states that can be conscious. One possible reply was to maintain that ordinary computers don't have sensory modalities and therefore, they do not generate the appropriate kind of fine-grained representations with mind-to-world direction of fit. This allowed the HOR theorist to rule out ordinary computers as being conscious, though it did require admitting that appropriately modified ones (sentient robots) could be. Beyond being adapted or tuned to other states of the system, Van Gulick does not set constraints on the kinds of FORs that can be conscious. All he says is that various lower level states already possess some degree of implicit meta-intentionality, prior to being integrated into a global state, but the range of possible states is extremely wide. This is due in part to the teleopragmatic view of the mind, which allows Van Gulick to conclude that such things implicitly represent (*e.g.,* bat's wings) or are implicitly self-aware (*e.g.,* the states that underwrite certain learning processes). On his view, it doesn't seem to matter whether the first-order states are appropriately fined-grained perceptual states or not. As long as they're representational it seems that almost anything goes, so long as it's tuned to some other feature of the organism's own cognitive system it will be representational. While his teleopragmatic view is focused on states and process which have naturally adapted to be tuned, there is nothing about the view that rules out similarly adapted artificial states from being appropriately tuned, and if so, then assuming some kind of computer could realize a global workspace (perhaps working memory), Van Gulick would have to concede that such a system would be conscious. We may count this as a strike against the

view, but I don't think Van Gulick would be troubled much by this, and he shouldn't be. His account already countenances an extremely broad notion of representation. From the perspective of the teleopragmatic theory it should not be surprising that certain kinds of artificial systems will be conscious. Again, this might seem counterintuitive from a common sense perspective, but it is hardly a fatal objection.

Regarding the threat from intramodular HORS, the global workspace is just that, it's a *global* workspace. It is not an encapsulated module. So the possibility of global workspace states that embody a heightened degree of reflexive self-awareness occurring *within* an encapsulated module does not threaten the HOGs view. We might envision a system w/multiple workspaces, but even if there were such a system, such spaces are importantly different from encapsulated modules.

The HOGs model also has a response to the third component of the too-easy problem (the argument from Freudian psychology). It is not clear that the Freudian cases *must* involve globally integrated states, *i.e.,* maybe they just involve ordinary HOTs. It seems that one mark of global accessibility is obviously missing in Freudian cases, *viz.,* such states are not available for verbal report. If one were to insist that the Freudian cases do involve globally integrated states that are severed between Freudian subjects, then the HOGS theorist can adopt the response already spelled out in Chapter 2. She can argue that the Freudian system already consists of two largely independent *subjects*. If both the conscious and the unconscious subjects traffic in global states, it wouldn't be surprising to find out that a globally integrated (conscious) state in one subject would be inaccessible to the other subject. After all, we've already admitted that there are two distinct *subjects* in one person, it wouldn't be surprising to find out the each subject has

its own distinct global workspace. All of this suggests that, while the too-easy objection forces the HOGS theorists to refine the view, the objection is by no means fatal. Even if the HOGS model can address the three main challenges for HOR theories, the model faces further difficulties.

Global workspace models can be cashed out in different ways. One way is to characterize the process in terms of information being globally *broadcast*. On this construal, information is actually transmitted to the (higher-order conferring) systems down-stream. Another way is to characterize the process in terms of *availability* or *accessibility*. On this construal, information may be transmitted to the global workspace, but not necessarily to downstream systems. Being in the global workspace, on this view, makes the information available to downstream systems. It's not clear which understanding Van Gulick is working with. He regularly says that first-order states are "recruited" into the global workspace, but that alone does not tell us whether or not they are then actually broadcast, or if in being recruited they are merely made available. Elsewhere I have argued that Van Gulick's view should be characterized in the latter way and that the view then results in a dispositional characterization of consciousness (Picciuto, 2011), which others have argued is problematic (Rosenthal, 2002b; Jehle and Kriegel, 2004; Kriegel 2005; and Weisberg, 2008). I no longer think that it's obvious Van Gulick's view should be taken this way. However, I don't know how to take it. It isn't clear what he means by being "recruited." A FORs being recruited is consistent with both of the above ways of characterizing the global workspace. If Van Gulick has the former in mind my earlier objection will apply, but if he has the later in mind it will not. Whatever Van Gulick himself actually has in mind, one might want to modify the view

endorsing the former as the better development of the view, and that would seem to evade the objection that the HOGS model is a dispositional account of consciousness.

All told, the HOGs model adequately addresses the rock problem, higher-order misrepresentation, and many, if not all, facets of the too-easy problem. Nevertheless, one might still wonder how the components of the model avoid these objections while explaining the distinctive features of phenomenal consciousness itself. In particular, no matter which sense of "recruitment" is appropriate for the view, how exactly does the fact that a mental state possessing implicit self-awareness gets recruited (either being delivered or being merely available) explain conscious experience? There is no clear answer provided by Van Gulick. Moreover, it is difficult to interpret what his actual goal is. Sometimes he seems to intend the view to explain some other kind of consciousness, but not phenomenal consciousness. Other times he suggests that he is, in fact, out to explain phenomenal experience, *e.g.,*

> [A] dark blue paperweight is present to me as part of my world, i.e., as part of the world that is present from my point of view, which is in turn as self defined by its location in that world of objects and appearance. That sort of implicit reference to the self is an essential component of *phenomenal content,* if not of intentional content in general. It is part of what distinguishes my *experiencing* the paperweight from merely representing it (2004, 85).

But again, just because a state of the system is widely broadcast, or made widely available to downstream systems does not tell us anything at all about why a state would be like anything for its subject. It may be that these conditions are coextensive with phenomenally conscious experience, but they do not *explain* why a state meeting them would be phenomenally conscious. Perhaps in being globally broadcast, the relevant state acquires additional content, or is made available to a distinctive system with a distinctive role, but that is not what Van Gulick says. If anything, the HOGS model seems merely to

rehash an account of something like access consciousness, tacking on a high degree of implicit metaintentionality. Thus, in spite of the virtues of the HOGS model, it winds up losing its account of consciousness itself, which is the primary goal of the model in the first place.

1.3 Carruthers Dual-Content Theory

According to Carruthers' "dual-content" theory, a mental state M is phenomenally conscious if and only if M is available to a higher-order thought-producing faculty (on Carruthers' view, the mindreading system). In virtue of M's availability to such a faculty, M acquires a higher-order content, which is integrated into M. Consequently, M both represents the world, via its first-order component, and itself, via its higher-order component.

Consider a simple example of a first-order nonconscious perception of red, the content of which can be symbolized as "[red]." When that red-representing state (call it "M") is made available to the mindreading system, which is capable of producing higher-order thoughts, M acquires a higher-order content, something like [experience of red] or [seems red]. Now M has two contents: {[red], [seems red]}. The "red" component represents some feature of the subject's environment. The "seems red" component represents the experience of red.

In spite of the clear self-representational components of Carruthers' view, it is more typical to find the view characterized as a version of dispositional HOT theory. It is not widely discussed explicitly as a version of self-representational theory.[27] But the self-

[27] See for example Weisberg (2008) and Kriegel (2004), but Cf. Kreigel (2006, 2009).

representational features of the view are evident and one of its central features, at least going back to Carruthers (2000).[28] For example, Carruthers writes,

> Where before these were first-order analog representations of the environment (and body), following the attachment of a HOT system these events take on an enriched dual content. Each experience of the world-body becomes at the same time a representation that just such an experience is taking place…But there don't actually need to be two physically distinct sets of representations to carry the two sets of perceptual contents [the first-order and higher-order components]…Rather, dual content comes for free with the availability of perceptual contents to the mind-reading faculty, or with the availability of those contents to HOT (2000, 242-43).

More recently, Carruthers writes that the dual-content of conscious states

> Immediately gives us an account of the key subjectivity, or "what-it-is-likeness," of phenomenally conscious experience. For by virtue of possessing a dual analog content, those experiences will acquire a subjective aspect…Hence they come to present *themselves* to us, as well as presenting properties of the world (or of the body) represented (2005, 107).

And,

> A number of people have proposed that conscious states are states that possess both first-order and higher-order content, presenting *themselves* to us as well as presenting some aspect of the world…My distinctive contribution has been to advance a naturalistic explanation of *how* one and the same state can come to possess both a first-order and a higher-order (self-referential) content (*ibid.* 9, fn. 7).

It should be evident from the above quotations that Carruthers' view is properly situated within the self-representational (or intrinsic higher-order) framework.

One might still be tempted to worry that it is the first-order state's availability to a numerically distinct HOT, or HOT-producing faculty, by which M acquires its dual content, but the worry is misguided. Being available to a HOT-producing faculty is a causal explanation for how M *becomes* appropriately self-representational. Existing

---

[28] It is also clear that Carruthers takes his view to be a self-representational view. See Carruthers (2012).

intrinsic theorists each offer different accounts of the process of integration (and I will offer my own in the next chapter). What unites the views as intrinsic HOR theories is not the causal processes they each postulate, which are supposed to lead to a conscious state being self-representational. Rather, what unites the views is that being a conscious state *consists in* its being appropriately self-representational.

One thing worth pointing out, though, is that while higher-order contents are integrated into the original lower-order state M, according to Carruthers, the higher-order content is represented *implicitly*, simply in virtue of the availability of M to the relevant downstream consumer system, in particular, the mindreading system. There is no explicit (structured) higher-order representational content that actually gets integrated into M. For Carruthers, the only sense in which the two contents are "integrated" is that they are possessed by the numerically same state. Availability is supposed to be what enables M to acquire higher-order content on the assumption of some version of consumer semantics, according to which for a state to have content is at least in part a matter of how the relevant downstream cognitive systems, or "consumers" might use that state. One significant way the mindreading system might consume the state is to think about it. Those who reject the account of content which is purported to account for M's dual content, may want also to claim that the view isn't truly self-representational, since in the absence of an adequate theory of content the view doesn't adequately account for the dual content of M (more on this below).

One strength of Carruthers' view as compared to Van Gulick's view is that Carruthers' view is independent of the global workspace theory. It is true that Carruthers *in fact* endorses (or is at least sympathetic to) the global workspace theory, but it isn't

really in virtue of the global broadcast per se that a conscious state is conscious. What makes a state M conscious is its availability to the mindreading system. Carruthers, does think that happens via global broadcasting, however, it doesn't have to. If the global workspace theory turns out to be false, Carruthers' account maintains its plausibility. For there would still have to be some way for the mindreading system to gain access to first-order states.[29] Van Gulick, on the other hand, thinks that conscious states are *partly constituted* by global accessibility. If the global workspace model fails, Van Gulick will need some other way to complete his account.

It is fairly obvious that, like Van Gulick's account, Carruthers' rules out the possibility of conscious rocks.[30] The notion of an object being available to the mindreading faculty is implausible. Is every rock on the planet "available" or only those in my immediate vicinity. Is the rock that is most in the focus of my visual field available if my eyes are closed? It is difficult to make sense out of availability here, without an analogous "workspace" in the environment that is inhabited by rocks. Also, setting aside the dispositional nature of availability, even when you occurrently think about a rock, you have a thought about it, but the rock isn't "consumed" by the mindreading faculty, and consequently, you don't bestow higher-order content upon the rock itself.[31]

---

[29] The question "which kind of first-order states?" is one that I will not engage with here. I am assuming at least that the HOT producing faculty has access to first-order perceptual states. As stated in Chapters 1 and 2, for the purposes of this dissertation I will set aside the question whether or not it also has access to attitudes.

[30] I already mentioned how Carruthers handles the too-easy problem in chapter 2, so I will not review the response here.

[31] Proponents of extended mind theories may not want to go along with the thoughts of this paragraph, but I'm not going to deal with the possibility of extended minds here.

Finally, since phenomenal character is parasitic on qualitative character (in that the very qualitative content of the target low-order representation M itself comprises one of the dual-contents that render M conscious, in virtue of the truth of some version of consumer semantics), Carruthers' account seems to rule out higher-order misrepresentation.

One potential weakness of the view is that, as with the HOGS model, it invokes the notion of implicit representation. Indeed the truth or falsity of the view hangs on whether or not implicit representation (via consumer semantics) can do the work it is intended to do. The higher-order content that is "integrated" into the first-order state M in virtue of being available to the mindreading faculty is implicitly represented. In an important sense, nothing actually happens to M intrinsically. M's change is dispositional. It becomes available for further use by the system, and thereby, is supposed to represent implicitly the first-order state itself. [32] For example, Jehle and Kriegel (2004) argue that it's implausible to think that categorical properties can be grounded by dispositional ones. But this is too strong.

The same concern about the appearance of consciousness as a categorical property that was raised above about the HOGS model may be raised against Carruthers' view. In direct response to Carruthers' view, Jehle and Kriegel (2004) assert that consciousness is a categorical property and that grounding consciousness, a categorical

---

Even if the mind is "extended" in some sense, it is hard to see how a higher-order state could bestow content upon a rock.

[32] Rosenthal (2002b), Jehle and Kriegel (2004), Kriegel (2005), and Weisberg (2008) have also made this point in their own ways, and it can be applied to other dispositional views, such as the PANIC theory discussed in chapter 1. This is a point to which I will return in section 2.4 when discussing Carruthers' dispositional dual-content theory. One might raise the same concern about Carruthers' view.

property, in a dispositional one as Carruthers does is implausible. But why should we think this is true? The only available data in support of the claim are phenomenological; presently, the only way we can assess conscious states is from the first-person perspective via introspection. One must undergo an occurrent state to know what it's like. But of course the property of being conscious will seem categorical when someone is actively undergoing an occurrent conscious experience. It's hard to see how it could not. It may be even harder to determine what theoretical force this appearance has; the phenomenology might not deliver a reliable verdict either way. As argued in the last section, the best that can be said is that consciousness *seems* like a categorical property. Our experience may represent our conscious states as being categorically conscious when in fact their underlying structures are dispositional.

Thus, again, one way to recast the objection against the dispositional element of the view is as a requirement that the dual-content theory be able to explain the *seeming* categorical nature of phenomenal conscious experience. As it stands, the dual-content theory does not explain this mental appearance. Carruthers' dual-content theory might well have the resources to do so, but I won't pursue that option here. Given the view's commitment to a specific theory of content, it is at least useful to consider alternatives which may have more minimal commitments about content and that avoid the categorical/dispositional issue altogether.[33]

---

[33] Notice that Jehle and Kriegel (2004) argue that the commitment to a specific theory of content, in particular commitment to consumer semantics renders the Carruthers' view implausible. This is far too strong. Since the theories under discussion are representational theories it is no surprise to find certain accounts committed to a specific theory of content (Van Gulick's too depends on the truth of his own teleopragmatic view). Here is a more charitable way to frame the worry. The main reason being committed to a

In defense of the dispositional element of the view, the HOGS theorist might argue that just because conscious experience seems categorical, that doesn't entail that the properties of the underlying structures must in fact be categorical. Just as consciousness might seem noninferential, even if it is underwritten by inferences of which the subject is not aware. The difference between the two cases is that appealing to nonconscious inference provides an explanation for the appearance of immediacy (if conscious states do not seem mediated by inferences, this will make them seem unmediated, or immediate). On the other hand, appealing to a dispositional property to explain the appearance of a categorical one does not explain why experience seems categorical, even if it is in fact underwritten by dispositional properties. So whether or not it really is implausible that categorical properties can be grounded in dispositional ones, we can sidestep that issue and recast the worry in terms of appearance. Our awareness of our own conscious states (their consciousness) certainly *seems* like a categorical feature. The HOGS model (or any dispositionalist account of consciousness) must explain this mental appearance. It must explain why our experience seems categorical, or how it possibly could, when it really isn't, *i.e.*, when in fact it is grounded by dispositional properties. The HOGS model, though, does no such thing, and merely asserting that conscious states

---

specific theory of content is worrisome is that no theory of content is flawless, and there is no consensus about which theory is best. Given the contentiousness involved with *all* theories of content, it should be considered a virtue of a theory of consciousness if it is neutral among the various theories of content to the greatest extent possible that one can be neutral. But surely a representational theory will have to assume at least something about the nature of content, even if it is quite general. Kriegel himself spells out his own view in terms of narrow contents, so he too is committed to at least some claims about the nature of mental content.

have a high degree of implicit meta-intentionality is not enough. We need an explanation for why that, in particular, would make consciousness seem categorical.

1.4 Cross-Order Information Integration

According to Kriegel's "cross-order information integration" (COII) account, a conscious state is comprised of three critical elements: i) a first-order representation, (ii) a higher-order representation, and (iii) a relationship of "cognitive unity" between (i) and (ii) (Kriegel, 2009, 233). Notice that (i) and (ii) are commitments of EHOR theorists as well. (iii) is what is supposed to make the view distinctively self-representational. Also, (i) determines the qualitative character of the state and (ii) determines what Kriegel calls the "subjective character" of the state. When the two components are integrated, the resulting complex state is phenomenally conscious; it has phenomenal character. For example, when I consciously experience the blue sky, there is a "bluish way it is like for me" (Kriegel, 2009, 1). (i) determines the bluish way (the qualitative character) and (ii) determines the "for me-ness" (the subjective character). When these two elements are integrated there is a for me-ness of the bluish way.

Conscious states are complexes because they enfold together an outwardly directed awareness of the world and an inwardly directed awareness of themselves. Kriegel writes,

> this sort of representational content may be produced simply through the integration of information carried by what are *initially separate* mental states. When the contents of these separate mental states are appropriately integrated, a (single) mental state arises which has just the right sort of representational content…it folds within it a representation of an external object and a representation of that representation (Kriegel, 2005, p. 46, my emphasis).

The two initially separate states that become integrated are "logical" parts of the unified conscious complex, the result of which is a single complex mental state that constitutes the state's being conscious. Without the relevant parts appropriately integrated, there simply is no conscious state. Kriegel does not tell us exactly what makes the integration appropriate. However, at least part of what appropriate integration consists in, he says, is being "psychologically real." I take it that this is what he means by "cognitive unity" in (iii) and that Kriegel uses the two expressions interchangeably.[34]

One subtle point of the COII hypothesis is Kriegel's gloss on the distinction between conscious and nonconscious perception. As we have seen, he argues that the former involve first-order qualitative parts integrated with higher-order parts. Non-conscious perceptions, on the other hand, do *not* have qualitative character. They have what Kriegel calls "schmalitative" character. Qualitative character and schmalitative character are both first-order. It's just that schmalitative states are never parts of conscious states, whereas qualitative states always are. This distinction is supposed to flow somehow from Kriegel's notion of a "constituting representation." Kriegel writes,

> To make sense of this, we must distinguish two kinds of property—the properties represented and the properties constituted by the representation. We may call the former schmalitative properties and the latter qualitative properties. Just as there is a distinction between being 6 foot tall and being represented to be 6 foot tall, so there is a difference between being schmalitatively bluish and being represented to be schmalitatively bluish…What I am proposing is that being qualitatively bluish ought to be identified with being represented to be schmalitatively bluish, (2009, 109-10).

[34] It would be helpful to have a better grasp of what Kriegel means by "psychologically real." Jehle and Kriegel (2004) maintain that a psychologically real process cannot be merely dispositional. It must amount to something "actually happening" within the subject, and it must be "temporally thick" (471).

The contents of schmalitative states are individuated narrowly. They consist of response-dependent appearance properties. Qualitative states are representations of response-dependent appearance properties.[35] However, it should be noted that they are not representations of *states* as representing those properties, *i.e.,* they are not metarepresentational.

To shed light on the distinction, consider the following progression, which is one way to understand the causal path that Kriegel has in mind. Suppose that at t_1 I see a tree, *i.e.,* at t_1 I undergo a schmalitative state. Call this state M_S. On Kriegel's view, the content of M_S is narrow. It consists of response-dependent appearance properties. In this case they are tree appearance properties. At t_2 M_S (or something else) causes my conscious state M_C, which is comprised of the higher-order subjective component M_H and the first-order qualitative component M_Q., *.i.e.,* M_C just is M_H and M_Q. Importantly, M_H represents M_Q, not M_s. Moreover, M_Q does not represent *the state* M_S. Rather, it represents the content of M_S, or as Kriegel puts it, the content of M_Q is constituted by the properties represented in M_S. In other words, my schmalitative state represents the appearance of a tree. When a conscious experience of the tree arises, the first-order component of the conscious experience is qualitatively tree-ish, because, on Kriegel's view, it represents the properties represented in M_S. Schmalitative states and qualitative states are numerically distinct states. And, qualitative states *are always parts of conscious states*. That is, the higher-order component always refers to the first-order qualitative state of which it is partly comprised, not to the first-order schmalitative state.

[35] Kriegel thinks that these are long disjunctions of neurophysiological states, but whether or not that is true doesn't matter for the present discussion.

Kriegel attempts to further develop his account by hypothesizing that the widely discussed "binding process" is one plausible way to account for the integration of first-order and higher-order states. As a first approximation, the binding problem is the problem of explaining how the brain binds together the relevant features of perception of a single object or event.[36] For example, shapes, colors, and motion are each detected and represented in different areas of the brain. But when, say, on the tennis court someone sees a roundish yellowish object moving toward her side of the net, she perceives this event as one cohesive event, not as several things coming at her at once, but independently of one another.

The specific theory of binding he puts to work is von der Malsburg's (1981) "binding-by-neural synchrony" view.[37] According to that view, when distinct populations of neurons in structurally distinct parts of the brain fire within milliseconds of each other, their content is bound together into what *seems* to be a single cohesive event. According to von der Malsburg's view, nearly synchronous firing in time represents the cohesion of the features represented. They are represented as being features of the same object. So to explain the bound features of the approaching tennis ball, the realizers of the representations of roundish, yellowish, and motion toward, fire nearly synchronously in time, and thus, are represented as being of the same object.

---

[36] This is a first approximation because many theorists distinguish between different aspects of binding. Such theorists think there are multiple binding problems. For example, Treisman (1999) dissects the binding problem into three separable problems (parsing, encoding, and structural description). Others deny there is a distinctive problem altogether (Hardcastle, 2008).

[37] For an earlier take on the neural synchrony account of binding, see Milner (1974).

Kriegel puts this account of binding to work to explain how the brain might realize the cross-order integration of mental states: just as there is a binding process that binds various environmental features of perceived objects, represented by distinct parts of the brain, into one cohesive perception, so too might there be a process (or an extension of the same process) that binds together first-order states with higher-order states into one cohesive conscious mental state. Information can be bound "across orders." Kriegel acknowledges that his appeal to binding is just a hypothesis (it is just one possible neural realization of his account), but he does think it provides some empirical basis for the cross-order integration that his account requires, even though the account doesn't rely on it. Nevertheless, the appeal to binding is supposed to do some explanatory work regarding the integration across orders of state.

Notice too that the standard binding problem deals with represented properties of objects, and how the different features of a single percept get bound together (the yellowness, roundness, and motion of the tennis, say) but Kriegel's implementation of binding involves binding between states, which presumably operate at a different level of abstraction. One might think that operating on a complex mental state as a whole, one with multiple contents and possibly attitudes themselves, is importantly different than binding together the represented properties of objects (color, shape, &c.).[38]

In Chapter 2, and above in this section, I claimed that higher-order misrepresentation is one of the motivating problems for IHOR theory. Even if one rejects that higher-order misrepresentation in fact provides adequate motivation for IHOR views, it is one factor that IHOR theorists themselves claim motivates their view. Gennaro, Van

[38] This point is stressed by Weisberg (2008, 170).

Gulick, Carruthers, and Kriegel all cite EHOR's inability to handle higher-order misrepresentation as motivation for their views. So it is clear that Kriegel takes the account to have an adequate way to deal with higher-order misrepresentation. The problem is that it doesn't.

First, if the binding of states that Kriegel has in mind is modeled on the first-order binding in perception, there is little reason to think that it would rule out higher-order misrepresentation. What Kriegel refers to as "the binding process" can itself go wrong. For example, there are various instances of illusory conjunctions, wherein the wrong features get bound to an object. There are different models that attempt to capture misbinding (*e.g.,* Treisman and Schmidt, 1982; Prinzmetal and Keysar, 1989), however they don't deny that misbinding actually occurs. Rather, they disagree about how to explain its occurrence. If misbinding occurs in first-order perception, we need some reason to think mental states cannot be misbound, and that it wouldn't occur at the higher-order level.

This is similar to discussions of (first-order) misrepresentation that one finds in the literature of mental content. Since we know that first-order misrepresentation occurs, an adequate theory of representation must accommodate and/or explain it. Similarly, the above mentioned binding-theorists all acknowledge that misbinding occurs, and that it must be accounted for, but not ruled out. On the other hand, higher-order misrepresentation is, at best, a theoretical posit. There is no data to support the claim that we actually undergo higher-order misrepresentations. Consequently, there is not the same kind of demand for a higher-order theory to accommodate it, but rather, given the problems it generates for the theory, the opposite is the case. Given the other constraints

of higher-order theories, there is more of a demand to explain how and why it cannot occur.

Setting that issue aside, near synchrony isn't synchrony. That means that the above progression from M_S to M_C isn't entirely accurately. Remember, I described the progression as beginning with M_S at t_1 and proceeding to M_C (the complex M_Q and M_H) at t_2. The question is: which two states are getting bound together? Since the resulting conscious state M_C is constituted by the complex M_Q and $M_{H,}$ it seems that they should be the two states that get bound together. However, the binding theory that Kriegel appeals to explains feature binding in terms of the *nearly* synchronous firing of *initially distinct* representations of features. To apply the view at the level of states, two things must be true. 1) the relevant states must be initially distinct states and 2) one state must preexist the other (admittedly this preexistence is extremely short lived, presumably in the millisecond range; but still, preexistence is preexistence). That means that 1) M_Q and M_H should be initially distinct states, and 2) that either M_Q exists prior to M_H or that M_H exists prior to M_Q. Now, it would be weird if the higher-order state were to preexist the first-order state. That's not an argument, though. Maybe it is that way and it's just an odd fact about consciousness. I don't think it matters either way for my point to be appreciated, though, so I will only consider the reverse scenario (that M_Q exists prior to M_H), which is the standard way to think about the causal path from a nonconscious state to a conscious one.

The problem is that, according to the schmalitative/qualitative distinction, M_Q can't exist prior to M_H. Indeed, neither M_Q nor M_H can precede or succeed the other. According to that distinction, M_Q can only exist *as part of a conscious state*. So M_Q and

M_H would have to come into existence as parts of the same state from the very beginning of their existence as M_C; they would have to arise in synchrony, not near synchrony, and that is in conflict with the characterization of binding to which Kriegel appeals. If M_Q doesn't exist prior to M_H, then there are no two initially distinct states to be bound together by near synchronous activation in the first place.

One might think that it is M_S rather than M_Q that gets bound to M_H, thereby "transforming" M_S, a schmalitative state, into M_Q, a qualitative one. But now one should wonder why being bound to M_H would matter to the character of M_S itself. The question is, why would being bound to M_H "transform" M_S from merely schmalitative to qualitative? If M_S wasn't qualitative before, why would it be now merely in virtue of being bound to a higher-order state? If the content of M_S consists of response-dependent properties before, why would it become a representation *of* those response dependent properties merely in virtue of binding? The claim that it would makes binding a mysterious process with powers that remain unexplained by the COII account. Moreover, M_S and M_Q are supposed to be coexisting numerically distinct states. As far as I can tell Kriegel intends it to be the case that M_S retains some aspect of its identity but "transforms" into M_Q. How could it, if M_Q is supposed to be a representation of the content of M_S? Here is another way to think about it. Consider the single state M. At t1 M represents the appearance of a tree in virtue of its response-dependent appearance properties. Then, at t2 once it is part of a conscious state, M represents those appearance properties. Why should we think that M at t1 is the same state at t2? There is nothing about the content of M at the two stages which suggests it has "transformed." They are numerically distinct states

with distinct contents. This is indicative of a larger problem the view has individuating states (Van Gulick, 2010; Weisberg, 2008), but I won't discuss that here.

The long and short of it is that cross-order binding and the qualitative/schmalitative distinction are at odds. Even if binding is left out the picture, the qualitative/schmalitative distinction is ill-motivated. It's not the slightest bit clear that the distinction is anything more than conceptual. Whenever a first-order state isn't part of a conscious state we call its character "schmalitative," but whenever the properties represented by that state are represented as part of a conscious state we call its character "qualitative." For what reason, other than the elaborate framework's ability to rule out higher-order misrepresentation, should we acknowledge this extra "schmalitative" layer? It adds little to the explanation of phenomenal consciousness, and there is no independent reason to think that the distinction is real.

The notion of a constituting representation is intriguing, and in Chapter 4 I will propose a way to make it more precise, but the qualitative/schmalitative distinction is not the best way to cash it out. Thus, even if the qualitative-schmalitative distinction were somehow amended to enable the COII account to avoid higher-order misrepresentation problems, it would do so at the cost of complicating the view well-beyond necessity. With no independent reason for acknowledging the distinction as anything more than conceptual, we can, and should, do without it.

2. Conclusion

That concludes my discussion of existing IHOR theories. Each of the four theories inherits the explanatory power of traditional EHOR theory. Additionally, each

has its own specific insights. Gennaro's WIV has us consider the scope of the conscious state to be wider than one might have thought, introducing a naturalistically kosher sense of "intrinsicality." Van Gulick's global availability highlights the role of cerebral celebrity and the wide influence conscious states do seem to have in cognition and in our daily lives. Carruthers' dual-content theory stresses the importance of integrating the first-order and higher-order components into a single state while advancing a naturalistic explanation for that process. And, Kriegel's COII account stresses the importance of integration being psychologically real as well as introducing the notion of a "constituting representation." Nevertheless, each view has its own pitfalls, and none of the existing views provides an entirely convincing account of the integration that is so critical to the intrinsic HOR model. Given the insights and pitfalls of existing IHOR views, it is useful to explore possible alternatives that draw from the strengths of existing views while leaving behind their apparent weaknesses. In the next chapter I lay the groundwork for one such alternative: what I call the quotational account of consciousness.

Chapter 4: The Mental Quotation Model of Consciousness

1. Introduction

This chapter begins to develop and defend what I will call the "mental quotation model of consciousness." According to the quotational model. A phenomenally conscious mental state is constituted by two main components: an appropriate first-order representation and a higher-order thought about that first-order representation, which represents the latter as a seeming its subject is currently undergoing. What makes the model different from other higher-order theories is that these are uniquely bound together by the operation of mental quotation. The structure of the higher-order thought is quotational. However, as I will explain below, while *mental* quotation is analogous to linguistic quotation in significant ways, it also importantly different. For example, the two may operate at different levels of abstraction. Linguistic quotation, on the one hand, does operate at the level of symbols (signs or expressions). Mental quotation, on the other hand, may operate at the level of "experiences," or contents.

In section 2 I distinguish between the original explanatory role of mental quotation in the phenomenal concept strategy and the explanatory role it plays in my account of consciousness. In section 3, I present the main characteristics of mental quotation. I draw both analogies and contrasts between linguistic and mental quotation. If the mental quotation model is accurate, it provides a simple and elegant way to explain many puzzles of conscious experience. There are several, and I cannot hope to address all of them here. In section 4, I illustrate how the quotational model can handle the core puzzling features introduced in Chapter 1 and the higher-order problems outlined in

Chapters 2 and 3. In particular, I argue that the model handles higher-order misrepresentation better than any existing view.

2. The Distinction Between the Model's Epistemological Precursor and the Present Metaphysical Model

The idea of mental quotation comes from the account of "phenomenal concepts" that David Papineau originated and once defended.[1] Papineau (2000) argued that we have unique concepts that we sometimes use to think about our own conscious states. These phenomenal concepts, he once argued, are best characterized as being closely analogous to linguistic quotation expressions. Their unique quotational structure played a crucial role in what has come to be known as "the phenomenal concept strategy."[2] The current received view of the phenomenal concept strategy is that its main purpose is not to explain what consciousness consists in, but to explain how we sometimes think about our own conscious states. Papineau's once-endorsed quotational version of the phenomenal concept strategy was, then, an epistemological project. My project in this chapter, on the other hand, is primarily metaphysical in the innocuous sense of explaining what

[1] The quotational account of phenomenal concepts is further developed by Balog (2012). Although he doesn't call his account "quotational," the idea of mental quotation is also employed by Block (2007), and it emerges in dualist form, but not as part of the phenomenal concept strategy, in Chalmers (2007).

[2] Papineau has abandoned the quotational account of phenomenal concepts. See Papineau (2007) for his more recent account. Also, there are other accounts of phenomenal concepts that have been employed in the strategy. Some alternative accounts are Loar's (1990), Carruthers' (2000), and Tye's (2000) recognitional accounts, and Perry's (2001) and O'Dea's (2002) indexical accounts of phenomenal concepts.

consciousness consists in, rather than merely explaining how we think about our own states that are already conscious by some independent factor.

Proponents of the phenomenal concept strategy attempt to undermine a well-known range of anti-physicalist arguments. While anti-physicalists point to mental phenomena that seem explicable only by dualism, phenomenal concept strategists explain them in a way that is consistent with physicalism. Thus, some or other physicalist account of consciousness remains possibly true. The apparent mystery of phenomenal consciousness, proponents of the strategy argue, is *merely* apparent. The anti-physcialist puzzles, they argue, are not puzzles arising from the nature of cosnciousness itself, but simply the distinctive way we sometimes think about our own conscious states. Proponents of the strategy do not argue that exercising phenomenal concepts is what makes a suitable mental state conscious. In contrast, the mental quotation model of consciousness that I am developing in this chapter does, in fact, claim that mentally quoting an appropriate lower-order state is what makes a conscious state conscious. Here I will set aside mental quotation as an instrument of the phenomenal concept strategy and put it to work in an account of what consciousness consists in. In Chapter 5 I will return later to discuss whether the positive account of consciousness in terms of mental quotation is also consistent with relying on mental quotation in the phenomenal concept strategy.

For Papineau (2000), mental quotations have two main components: a quoting component and an occurrent perceptual (or imagistic) state that is quoted. Papineau's original schematization of the structure is 'the experience _ _ _', where 'the experience' is an operator that operates over an occurrent conscious experience, which fills in the

blank.[3] On Papineau's view, "phenomenal concepts are compound terms, formed by entering some state of perceptual classification or recreation into the frame provided by a general experience operator…Very roughly speaking, we refer to a certain experience by producing an example of it" (116). While according to my own quotational model of consciousness, mental quotations quote perceptual states that would be *nonconscious* were it not for being quoted, this is the general schema that I will borrow and adapt for an account of phenomenal consciousness itself.

3. Characteristics of Mental Quotation

According to the mental quotation model of consciousness, mental quotations (and, thus, conscious states) have two core components: a higher-order quoting concept that takes an occurrent, appropriate first-order state (an "experience") as its object. While Papineau used 'the experience_ _ _' as the concept frame, 'experience' is heavily ambiguous, especially between conscious and non-conscious experience.[4] The model I am proposing is importantly different. On the current model, the quotational concept frame targets a perceptual state *that would otherwise be nonconscious, were it not for being quoted*, so a fully developed concept of conscious experience is not required. This

[3] Notice that Papineau's schematization is spelled out in terms of a definite description. This is potentially misleading. As far as I know, it should not be taken as an endorsement of the definite description theory of quotation (Tarski, 1956; Geach, 1957; Quine, 1960; Davidson, 1979). It is just a shorthand way of representing the structure for heuristic purposes.

[4] Admittedly, when one thinks of experience intuitively it is hard *not* to think of conscious experience. But the commonsense notions of experience may have little to do with the nature of the sorts of state underlying "experiences" that are relevant to a theory of consciousness.

difference suggests the following emendation to Papineau's original schema. In place of the conscious experience operator ('the experience _ _ _') we can analyze the first core component as a "seems operator:" 'SEEMS <__>'. This might appear to be a minor terminological change, however it actually bears on more fundamental issues. For example, the *seems* operator is more elementary than a full-blown experience operator would be, arguably arising earlier in evolutionary, and perhaps developmental, history. It requires a concept of the is/seems distinction, but not necessarily a concept of *conscious* experience, which might well arise from a more sophisticated faculty. It might be that the faculty that conceptualizes the is/seems distinction is an evolutionary or developmental precursor to a full-blown mindreading faculty, but if so, it is much less sophisticated. This is important because we should not rule out a priori species lower down the evolutionary scale (or creatures earlier in their development) from undergoing phenomenally conscious states, even if in the end it turns out we must concede that they do not undergo conscious states.[5]

The second core component of a mental quotation is the corresponding occurrent experience that is the object of the quotation. In contrast to Papineau's account, which characterized the second component as an occurrent conscious experience, on the quotational model of consciousness, the second component is an occurrent, appropriate

[5] There is at least some evidence that non-human animals have a concept of the is/seems distinction (Lurz, 2011). And there has been much debate about higher-order theories and the issues of animal and infant consciousness. I will not try to settle the debate in this dissertation, but the issues will be touched on in Chapter 5, when discussing some future directions for the quotational model.

first-order representation, *i.e.,* it is an occurrent *nonconscious* experience (if considered singly, independently of being a target of quotation).[6]

As alluded to in Chapter 1, here I want to highlight that different kinds of state can function as the occurrent "experience," *e.g.,* perceptual states, imaginative creations/recreations, episodic memories, illusions, and hallucinations. For the purposes of the mental quotation, there is nothing about the external world, regarding the content of the concept, that need obtain (*i.e.,* the internalism/externalism debate about mental content is orthogonal). Moreover, the account is intended to apply to a variety of *sensory* states, not only perceptual states. As outlined in Chapter 1, an "appropriate" first-order representation is a state *something like* the characteristic output of a prototypical sensory system. Most significantly, such states are mixed conceptual/non-conceptual fine-grained representations with mind-to world direction of fit (see pp. 34, 55-56).

The quoting feature of a mental quotation (the *seems* operator) is the concept frame. As mentioned above, it operates over occurrent sensory states (whether veridical or not). Together, the two components form a unique structure that constitutes the way things consciously seem. On its own, the concept frame can be characterized as being something like what Frege meant by 'unsaturated': it *requires* the second component (the occurrent experience) to be a complete mental quotation, *i.e.,* an occurrent experience is *mandatory* for completion of the quotational concept, and without one, the concept frame

[6] As I discussed in Chapter 2, it might be the case that, in addition to perceptual states, attitudes are also sometimes phenomenally conscious. Indeed, several theorists argue that they are and would consider such states appropriate first-order representations. In this dissertation I am setting aside attitude states and dealing only with perceptual states.

has no satisfaction conditions.[7] Importantly, the experience picked out by the quotational structure *partly constitutes* the complete quotational concept that picks it out. The mental quotation refers to the content it does partly in virtue of presenting an active token of it via one of its own parts.

Here is a very rough example of how the above components mesh (to be fleshed out over the next few sections). When I consciously perceive (or imagine, or remember, or hallucinate), say, a white swan on a dark pond, according to the view under discussion, I undergo an appropriate first-order representation. In this example, the appropriate first-order representation is a mixed conceptual/non-conceptual visual representation of (at least) the white swan on a dark pond. That representation is then delivered to a conceptual system that wields quotational concepts. This generates a metarepresentational state, *viz.* SEEMS <white swan on a dark pond>, that represents my white-swan-on-a-dark-pond-representing state *as* a white-swan-on-a-dark-pond representing state *that I am undergoing,* but which also represents the white swan on a dark pond (it displays or activates that very sensory white-swan-on-a-dark-pond-representing state). This is the sense in which the mental quotation does "double duty," to borrow Block's expression: it represents both whatever is represented by the first-order state as well as representing that first-order state itself. This marks one significant distinction between mental quotation and linguistic quotation as it is ordinarily employed. The distinction between the two will be further addressed in section 2.2.

[7] Cf. Frege (1892/1976; 1892a/1976a) and Recanati (2004) on saturation.

Notice that the conscious awareness involved in the above example is an awareness of the white-swan-representing state as an appearance, but not necessarily as a perception. In other words, I am not claiming that, on the quotational model, a subject is automatically aware of her perceptual state *as a perceptual* (as opposed to imagistic) *state*. Rather, she is merely aware of its seeming, fine-grained content. It would require a further judgment about that state to attempt to determine the specific sensory profile of the state, *e.g.,* whether it is a veridical perception or a hallucination.

Pulling all of this together, a conscious mental quotation is a higher-order representational thought that contains, as a constituent, the appropriate first-order representation that it targets; the first-order state is intrinsic to the higher-order state. The resulting concept is a concept that applies to an experience that would otherwise be non-conscious, rendering it conscious. In other words, the result is a kind of self-representational state. The quotational model is properly characterized as a kind of self-representational, or intrinsic higher-order, theory. In particular, because the mental quotation is a quotational *concept*, the model is a kind of intrinsic higher-order thought theory. That is the core structure of mental quotation. The quotational component of the view and its rough location within perception and cognition is further developed in the remainder of this chapter.

Before proceeding, two disclaimers are in order. The first is that the quotation process, as just described, does not require that the subject consciously think the thought "It seems to me," nor that the subject is cognitively sophisticated enough to articulate such a thought in inner-speech. It might well be the default assumption built into the mind-reading system (or its evolutionary precursors) that any incoming state is an

appearance state that is the subject's (or system's) own state. For competent thinkers, deploying mental quotations is effortless and automatic. As an analogy, consider the idea that children before a certain age have not yet developed a concept of belief, however, they are able to believe things; they have beliefs (if there are such things) running in their systems. Similarly, a subject might lack a concept of experience (whether conscious or not) and yet still mentally quote states that would be nonconscious were it not for their being quoted. In other words, a developed concept of experience is not required to target mentally a non-conscious experience via a quotational concept.[8]

The second disclaimer is that, while the *seems* operator is implicitly assertoric (it asserts that there is a seeming of *x* going on), the view does not simply repackage the old adverbial theory, *e.g., "x* seems greenly" (Sellars, 1963, 1967, 1971, 1975). For the state in question does not merely assert *that* there is a seeming *x*-ly underway. Rather, it implicitly asserts that a seeming is underway in virtue of employing the very state to which it refers, and it does so by presenting it as one of its own parts.

3.1 Mental Quotation vs. Linguistic Quotation

Linguistic quotation expressions exhibit many philosophically interesting features. There is much debate about how to characterize these features and isolate the ones that are essential. For example, some authors argue that quotation *marks* themselves are derivative, arising from quotational *usage*, and therefore, that they are eliminable (Davidson, 1969; Washington 1992; Saka, 1998, 1999, 2005; Reimer, 1996, 2005; and

[8] This raises an interesting question about how such concepts are acquired, since one point of the phenomenal concept strategy is that one cannot acquire phenomenal concepts of experience without having first experienced the relevant experience. I will return to this question is section 4.

Recanati, 2001). Others argue that quotation marks are a more fundamental semantic or syntactic feature of language (Cappelen and Lepore, 2006).

One particularly striking feature of quotation expressions is the apparent intimacy, or proximity, between the quoter and quotee. For example, it looks like the expression "Fodor" literally contains 'Fodor', its quotee and its semantic value. But 'Fodor', on the other hand, *does not* contain its semantic value (the man himself!). Different theorists of linguistic quotation have touched on this point in different ways. Quine wrote that "A quotation is…a *hieroglyph*…[that] designates its object…by picturing it (1940, 26). According to Davidson, "[A] quotation somehow pictures what it is about" (1979, 82). Bennett wrote that "[W]hat is displayed in a quotation is systematically related to what it names" (1988, 401). On Saka's view, "[W]e can go from knowing the quotation of any expression to knowing the expression itself" (1998, 116). And here are Cappelen and Lepore: the relationship between a quotation and its semantic value is a relationship "like no other kind of expression bears to its semantic values" (2007, 24-25), and "For any quotable item e, if a quotation expression Q quotes e, then e is contained in Q. Containment describes a basic feature of quotation expressions" (124-125). On their view, "quotation is a *sui generis* device for connecting language to the world…As such, quotation constitutes one of the most basic ways in which language connects to the world," (5). According to Cappelen and Lepore, quotation expressions literally contain their semantic values. I will not be endorsing Cappelen and Lepore's minimalist view, neither as an account of linguistic quotation nor as an account of mental quotation. However, some of the core features of their account do lend themselves to the way I am

understanding the process of *mental* quotation. This will come into view throughout the course of this section.

There are also thought to be different kinds of quotation, *e.g.,* pure, direct, indirect, and mixed. Here are some examples:

1) 'Fodor' refers to Fodor

2) 'New York' has seven letters.[9]

3) Fodor said, 'The semicolons aren't safe'.

4) Fodor said that the semicolons aren't safe.

5) Fodor said that the semicolons 'aren't safe'.

(1) and (2) are examples of what Cappelen and Lepore call "pure" quotation. As they describe it, "In pure quotation, there is no attribution to any utterance or saying event" (2007, 14). (3)-(5) respectively represent direct, indirect, and mixed quotation. According to Cappelen and Lepore (3) quotes by mentioning words that Fodor uttered. (4) quotes Fodor, but could be true even if he uttered none of those particular words, *e.g.,* if he said 'Those marks better watch out' or 'Look out, quasi periods'. (5) quotes by reporting what Fodor said but by attributing to him only an utterance of 'aren't safe'.

There is a sense in which the mental quotations that I have in mind are direct. For example, in SEEMS <Φ> there is no attribution to a thinking event (the putative analog of an utterance event). There is, however, an attribution of a seeming event, so there is also a sense in which mental quotations are indirect. The seems operator is implicitly

[9] I will not deal with scare quoting. I am concerned with the mental analogs of representations such as *'Chicago' has seven letters,* but not with the mental analogs of representations such as *This is "Chicago" style pizza.* (said with ridicule).

assertive in that it asserts a current seeming is underway. Ultimately there might not be a direct analogy between mental quotation and one of the above-mentioned kinds of linguistic quotation.

Linguistic quotation expressions involve some canonical sign system that includes a quotation symbol, *e.g.,* single apostrophes in British English, double apostrophes in American English, double angles in some European languages. There are in fact limitless ways that one can symbolize the operation of quotation. One question that immediately arises in response to the hypothesis that there are mental quotations is: *What does the quoting?* And if quoting involves referring, then *What does the referring?* This latter question is of interest to theorists of linguistic quotation as well. Cappelen and Lepore (2012) offer three "guiding questions" for a theory of (linguistic) quotation, the first of which is just that:

(Q1) What does the quoting/referring?

Their second two guiding questions are:

(Q2) To what does the relevant component refer?
(Q3) How does the relevant component quote/refer?

These questions are well-suited to guide a theory of *mental* quotation as well. They suggest the following analogous guiding questions for a theory of mental quotation.

(MQ1) What does the mental quoting/referring?
(MQ2) To what does the relevant mental component refer?
(MQ3) How does the relevant mental component quote/refer?

Let's look at each of these in turn.

MQ1: What does the mental quoting?

Notice that (Q1) asks which component of a quotation expression refers (if any does). One difference between linguistic quotation and mental quotation is that the very existence of quotation *marks* is not up for debate regarding linguistic quotation. While there is a debate about whether quotation marks are, on the one hand, essential to linguistic quotation or whether they are, on the other hand, derivative and eliminable, as use-theorists claim that they are, that debate is not about whether quotation marks can or cannot be used to indicate quotational usage. Moreover that debate is not about whether or not there really are such things as quotation marks in the first place. But a theory of mental quotation must answer exactly this more fundamental question first. Are there mental quotation marks in the first place? If not, then in what non-metaphorical sense are there mental *quotations*? Part of the answer will involve answering related questions: Is there some canonical symbolization system for mental quotation marks? Or are mental quotation marks not just eliminable like their linguistic counterparts might be (at least on some views), but altogether nonexistent? Might there be an analogous mental use-theory for mental quotation?

Here are five possible ways to characterize mental quotation *marks*. Mental quotation marks could be characterized as 1) explicit quotation symbols presumably symbolized in a language of thought, 2) imagistic mental pointers that require corresponding demonstrations, 3) causal "control structures," 4) entirely eliminable, in favor of a kind of mental use-theory, or 5) as concepts that quote abstractly by taking experiences as their objects.

(1) is how Papineau originally characterized mental quotations. As discussed above, Papineau thought that mental quotation marks were operators, presumably (but

not necessarily) symbolized in a language of thought. Something like THE EXPERIENCE<blank>, where 'THE EXPERIENCE' functions as the quotation operator. On this way of formulating mental quotation, a canonical symbolization convention would seem required in the language of thought.

I do not know of anybody who holds a view like (2), but it's worth considering, given that the operation of quotation strikes many as involving a demonstrative element (even if the characterization of quotation ultimately will not be in terms of demonstratives alone). Suppose one thinks that quotations have some kind of demonstrative element. Now, on Kaplan's very plausible account of demonstratives, a demonstrative on its own has no descriptive content. Demonstratives require roughly simultaneous "demonstrations" to determine their referents. Different things can serve as demonstrations. For example, uttering 'that' while pointing or uttering 'that balloon' while in the presence of some balloons. In the first case the pointing is the demonstration. In the latter case the word 'balloon' is. Mental quotation marks, then, can be characterized as imagistic mental demonstratives., *e.g.,* a kind of finger pointing in the head.

Another possibility is that mental quotation "marks" are not explicitly represented, but rather they are constrained by what Prinz (2007) calls "control structures." Prinz does not discuss mental quotation, but he has his own idea about control structures underwriting the process he calls "mental pointing." Since quotation involves, but is not exhausted by a kind of pointing, a similar sort of process could be put to work to help explain the pointing and presenting of which mental quotation consists, or if one thinks there is a demonstrative element, then it could be put to work to explain that component

of mental quotation. Roughly, according to Prinz a control structure has causal control over some range of entities (13). On Prinz's view, "[p]henomenal demonstratives use representations of objects in space to direct focal attention on a perceived scene. They are individuated by their causal powers" (13). Similarly, mental quotation marks range over (at least in part) perceptual contents and mental quotation expressions employ occurrent representations of, say, the focused objects in space *as a perceived scene*. Whether or not the resulting structure qualifies as a concept is an independent question. On Prinz's view the control structures that constitute mental pointing are explicitly not concepts. His purpose is to present an alternative to the phenomenal concept strategy.

Next, even if one holds a language of thought view of the mind and thinks that that mental quotation marks are explicitly represented in a language of thought, one might still think those quotation marks merely indicate quotational usage. Or more precisely in the case of mental quotation, they indicate quotational *intentions*, or the quotational operation. On such a view, mental quotation marks are analogously eliminable. More precisely, there is no need for elimination because there aren't any in the first place. There is only fundamentally quotational non-conscious intention.

Finally, mental quotations can be characterized abstractly, as involving concepts that take experiences as their objects. In that regard they function at the intentional level, *i.e.,* they function at the level of content, not at the symbolic level.

To make mental quotation plausible, one need only show that there is some possible plausible story about quoting in the head. At this point I am trying to show that mental quotation is non-metaphorically plausible. To do that I do not need to choose between these options. Whatever view is the best view is an empirical question, and there

is no evidence that favors one view of mental quotation marks over the others. However, there are some conceptual reasons that make some of the options more plausible than others.

Of the four options I find (1) and (5) the most plausible. First, consider (1). While the language of thought hypothesis is not universally accepted, it has many explanatory virtues, and if there is such a thing one might plausibly argue that it contains mental quotation marks as a primitive syntactic symbol. The resulting mental "expressions" would be structured much like their linguistic counterparts, but would involve the interface of the conceptual and the non-conceptual (insofar as 'nonconceptual' is characterized symbolically as well). Cashing out mental quotation marks this way does have consequences for the answers one ought to supply for MQ2 and MQ3, though. It would seem that such a view would need to be committed to a LOT account of sensation. Otherwise, the quotational procedure does become mysterious (but not obviously false). One might wonder how a quotation in the language of thought could quote a nonsymbolic entity.

On the other hand, (2) is less plausible for many of the same reasons that plague inner sense theories. For if mental quotation marks are imagistic pointers, there would need to be some inner perceptual mechanism to perceive (nonconsciously) the pointer-image. If not, then it's unclear what would make it the case that the pointer functions as the requisite "demonstration" in the relevant Kaplanian sense. But there is no good reason to think that there is any internal perceptual faculty in the first place, so I won't pursue this option any further.

(3) is intriguing and has the virtue of providing a more general characterization that is consistent with both language of thought style characterizations and non-language of thought style characterizations. But 3 would require much more elaboration, and I'm not going to provide that here.

(4) is also intriguing but less plausible as a direct analog of linguistic use-theories. It might sound odd to talk about mental quotation marks as "indicating" quotational usage, since there are no other competent speakers to interpret that usage, as there are in the case of presentations of linguistic quotation, but here is where the notion of constitution becomes important. Mental quotation marks characterized in terms of quotational intentions would *constitute* the quotation operation. One possibility is that a distinct variable would be co-opted in distinct contexts of thought, or activation, much like neurological structures can be co-opted in certain cases for distinct processes after one sustains injury.

In addition to (1), (5) has initial plausibility for the account of consciousness being developed in this chapter. Considered abstractly, mental quotation need only involve a concept that takes a perceptual state as its object (or quotee), thereby referring to it essentially by incorporating the perceptual state into the thought. On this way of cashing out the view, mental quotations are admittedly "quotations" in a weaker sense, perhaps, only by analogy. But for the purposes of an intrinsic higher-order thought theory of consciousness, the view need not be committed to the claim that the operation of mental quotation and linguistic quotation are isomorphic. Rather, the crucial point is that they both have quite similar results: both linguistic and mental quotations refer to a target

in virtue of having their targets embedded in the thoughts that quote them. This is the option that I will embrace for the purposes of an account of consciousness.

MQ2: How do mental quotations refer?

As discussed in section 2.1, for Papineau a mental quotation is a compound concept consisting of a concept frame and a conscious perceptual experience. Papineau explains how such a compound manages to refer by appeal to a version of teleosemantics. Mental quotations, on his view, derive their referential power from "facts about the causes or biological functions of the deployment of those terms" (2000, 116). Mental quotations (which, for him are phenomenal concepts) are compound referring terms composed of the experience operator and a perceptual filling (on his view, a conscious state). The referential power of the compound concept derives from the systematic contributions of its parts. In this case, the contribution of the parts to the semantic value of the compound depends on the systematic contributions those parts make to the causes or biological functions of the whole phenomenal concept. Papineau does not tell us much at all about what those systematic contributions are, but he does conjecture that the "biological purpose of the whole might be to enable our ancestors to better predict the behavior of others, enabling them to anticipate their own future experiences or to facilitate reflection on the epistemological credentials of their own beliefs and the beliefs of others" (116, fn 8).

Must we endorse this particular teleosemantic account to explain how mental quotations manage to refer? No. But surely it constitutes one possible account of how mental quotations could possibly refer. Whether or not teleosemantics is ultimately successful is a completely different issue.

According to Balog's account of mental quotation, quotational concepts manage to refer in virtue of their conceptual roles. She says the operation of linguistic quotation consists in a speaker's disposition to accept all iterations of a disquotation schema, *e.g.,* "q" refers to 'q', '"q"' refers to "q" and so on (forthcoming, 27-28). What accounts for the fact that quotation marks indicate the operation of quotation (in English) is that all competent speakers who are users of the marks and who understand the meaning of 'refers' are disposed to accept all instances of the above schema.

For Balog the mental quotation schema is similar. However, the sentences are expressed in Mentalese. For example, what accounts for the semantics of the operation of *mental* quotation is that competent *thinkers* who have a concept of reference are disposed to accept all instances of the *mental* disquotation schema. Balog maintains Papineau's original schema (THE EXPERIENCE <...>). Here, for consistency, I will substitute for Balog's the schema I suggested above. Her mental disquotation schema becomes

SEEMS<Φ> refers to Φ

SEEMS SEEMS<Φ> refers to SEEMS Φ

where SEEMS<Φ> ranges over token experiences, thereby referring to either a token or a type.

Mental quotation is like linguistic quotation, she writes, "with one important difference. The difference is that, unlike linguistic quotation, what is between the mental quotes…at the first level is not a mental word but a mental representation that is not itself a word; it is an experience" (23). Disquotation, then, is the conceptual role that determines the content of a token mental quotation. It has the content it does because the speaker is disposed to accept the disquotation schema.

The mere disposition to accept the disquotation schema leaves a lot unexplained about mental quotation. One might wonder about why the thinker is disposed to accept the disquotation schema for linguistic quotation. The answer with linguistic quotation seems clear enough: the speaker is disposed to accept it because of pragmatic conventions. On the other hand, in what sense could *the thinker* be disposed to accept the *mental* quotation schema? Balog assumes, for the most part, that mental quotation operates in the language of thought (even though it's consistent with other interpretations). How then do the mental analogs of the linguistic conventions arise in the language of thought? Balog doesn't consider this question. Thus, the core of the view is absent. My point here is not to provide a detailed assessment of Balog's account, though. I merely want to show that there is some or other way of accounting for the content of mental quotations. One can envision more robust conceptual roles that might determine the content of mental quotation. For example, Carruthers' dual-content theory is closely related. On his account a mental state M can acquire content by being available for further use to systems downstream. Part of what determines the content of M is how the downstream systems, or "consumers" will use the state. One can envision a similar conceptual role story to strengthen Balog's disquotation role. In fact, it may be that what Balog is really trying to get at is something like that story. For example, what makes it the case that a mental quotation has the content it does is just that downstream systems will use the state is certain ways, *e.g.,* they might disquote the state.

There may be other characteristic roles of quotation that could further support the view. Again, though, my aim here is to point out that, in addition to Papineau's teleosemantic account, there is another possible story to be told about how mental

quotations manage to refer. And here too, I'm not endorsing conceptual role semantics. Rather, I am acknowledging another possible explanation of a mechanism for the semantics of mental quotation. Whether or not conceptual role semantics succeeds in the end is a different story.

Balog explicitly claims that informational and nomological accounts require an "external" relation between a concept and its referent, which is "unlike constitution." This makes them apparently unsuitable candidates for an explanation of self-reference. On the contrary, I will spell out how a nomological account can do just as well as (or at least no worse than) teleosemantic and conceptual role accounts of mental quotation.

On an asymmetric dependency view of mental quotation (defended by no one, but plausible enough) the SEEMS concept frame would asymmetrically covary with the sensory contents that system charged with deploying quotational concepts encounters (including veridical perceptions, imaginative creations/recreations in memory, hallucinations, or whatever). For example, it may be that quotational concepts are deployed top-down, by the mindreading system. Following that line of thought, when the mindreading system "is in the presence" of a sensory content, the systems SEEMS concepts frame gets tokened (thereby embedding the sensory state, just as perceptual states embed concepts). Furthermore, initially such sensory contents are "external" to the concept frame. Thus, "in the presence" of an active perceptual state, my quotational SEEMS concept gets tokened, thereby embedding the very state itself. Continuing with this line of thought, erroneous cases, wherein the seeming is tokened in the absence of an actual perceptual content, depend on veridical instantiations, but on an asymmetric dependency view, the veridical instances don't depend on the erroneous ones.

I am not endorsing any of these as an account of the content of mental quotation. For the purpose of this dissertation I only need to show that there is at least one available account of the content of mental quotation. What I have in fact sketched is the possibility that (setting aside their deficiencies), not just one, but all three of the main existing theories of content might possibly have the resources to explain the content of mental quotation. Neutrality isn't merely an easy way out of this. It is, in fact, a virtue of the quotational model that the model is consistent with, and neutral about, all three main accounts of content. In fact, this is one of the main arguments against Carruthers dual-content theory that was discussed in Chapter 3, *viz.,* that if consumer semantics is false, it looks like the dual-content theory is false, or at least, would require extensive overhauling. It is also one argument against Van Gulick's HOGS model, which essentially hangs on his bizarre teleopragmatic theory of content. The quotational model has no such commitments, so the theory of consciousness it provides is not held hostage to any particular theory of content.

MQ3: To what do mental quotations refer?

What are the quotable items for a mental quotation? Are brain states literally nested? Must the neural realizers of mental quotation be nested themselves? Is a symbol in a language of thought what is quoted, or a symbol *and* its semantic values? Perceptions presumably include a non-conceptual component. How are non-conceptual components quoted? These are all questions about which a theory of mental quotation must have something to say.

I think it is clear from what has just been discussed that the main proponents of mental quotation maintain that mental quotations quote at the level of content. They do not merely quote symbols. They do not quote not neural realizers. Papineau and Balog couch their characterization of the quotable items in terms of (conscious) experiences. Similarly, on the quotational account of consciousness being developed here, the critical quotable items are experiences. Importantly, though, I am using a deflated notion of experience. For one thing, as explained in Chapter 2, I am only dealing with sensation. Whether or not there are cognitive "experiences," or a phenomenology of attitudinal states is something I set aside. Second, I reject Block's claim that phenomenal consciousness just is experience. As many others do, I am working with a notion of 'experience' that includes both conscious and non-conscious experience. All I take a non-conscious perceptual experience to be is a sensory representation (a mixed conceptual/non-conceptual representation with mind-to-world direction of fit, as outlined in Chapter 1). These are typically the outputs of a prototypical sensory system, but in principle they wouldn't have to be.

Another contrast with Papineau's and Balog's accounts is that the mental quotation account of consciousness is perfectly at home with experiences construed as LOT terms or expressions, *e.g.,* restricted predicates, or "sensational sentences."[10] To sum up, there are several options for making mental quotation plausible.

[10] See Rey (1992, 1993, 1998).

3.2 Containment

The apparent intimacy between the quoter and quotee is of interest both to theorists of linguistic and mental quotation. It is the main feature that has attracted mental quotation theorists in the phenomenal concepts literature. Quotation expressions seem to *contain* (or are partly *constituted* by) their semantic values in a way that ordinary referring expressions do not. Analogously, mental quotations (quotational concepts) seem to contain or are partly constituted by the *experiences* to which they refer in a way that non-quotational concepts do not.

One issue related to the idea that quotations involve containment or constituting representations is whether quotations require spatial containment in particular. Tick marks, our predominate symbolization convention for linguistic quotation, certainly employ spatial containment. It might be thought that, if quotation in general requires spatial containment, that constitutes an obvious-seeming objection to mental quotation. For one might plausibly wonder how one mental state (or brain state) could literally spatially contain another. And, if it cannot, then mental quotation is a nonstarter.

In fact, there has been some discussion recently about nested neural states, which may or may not make plausible the idea of one brain state containing another. According to Feinberg's (2000, 2001, 2009) "nested hierarchy theory of consciousness," lower levels of the brain are "nested" within higher-levels. As we ascend through the hierarchy, the level of complexity increases. When a conscious state occurs, the lower-level features of, say, a perception are activated as parts of the higher-order features.

The nested neural hierarchy theory might lend some amount of plausibility to the quotational account insofar as it describes a possible neural realization. However, if it can

be shown that spatial containment isn't a necessary condition of quotation in the first place, then the issue of spatial containment can be set aside, and the appeal to nested neural hierarchies can be postponed until more conclusive evidence is available, and until more developed theories are constructed. This is the route that I will take. I will illustrate the point that quotation in general does not require spatial containment.[11]

Our current convention is to symbolize quotation using tick marks that surround the quoted material. However, that is merely what our own symbolization convention happens to be. To think that the operation of quotation requires spatial containment involves confusion between how we symbolize the quotation operation and the quotation operation itself.

Consider our canonical sign system and the expression "Lepore." Presumably "Lepore" quotes 'Lepore' ('Lepore' falls between the outer tick marks of "Lepore." "Lepore" *spatially contains* 'Lepore'). It cannot be the case that "Lepore" quotes 'Washington', *i.e.,* "Lepore" must quote 'Lepore' and not 'Washington' because 'Washington' cannot fall within the outer tick marks of "Lepore". "Lepore" is partly constituted by 'Lepore'.

The issue is not whether "Lepore" can quote an expression ('Lepore') that means, names, or refers to something other than, say, *the philosopher of language at Rutgers*, say, the *Capital City of the United States.* Rather the issue is whether "Lepore" can quote

[11] Everyone allows that there are countless ways to symbolize quotation, not all of which employ spatial containment. Oddly, this is something with which everyone agrees, that is, until they come to consider *mental* quotational.

anything other than its constituent sign or expression.[12] Because the quotation expression "Lepore" is partly constituted by 'Lepore', the answer seems to be that it cannot.[13] The fact that "Lepore" spatially contains 'Lepore' is determined by a symbolization convention. But the *operation* symbolized by that convention is more primitive and could be captured in any number of ways, *e.g.,* by prefixing each quoted word with '@', as in '@The @semicolons @aren't @safe'. Nested quotes could then be expressed by appending '@'s, as in '@Fodor @said, @@The @@semicolons @@aren't @@safe'. Prefixing the words quoted with '@'s is clunky, but I suspect nothing substantive would be lost from doing so. Moreover, if '@'s are too clunky, there are other ways to symbolize the operation of quotation, *e.g.,* italics, or the prefix 'quote/unquote', both of which are, in fact, used interchangeably with tick marks. Thus, spatial containment is not a necessary condition for quotation, *i.e.,* not even for *linguistic* quotation. If so, it is plausible to conclude that the issue of spatially containment is not a problem for the idea of mental quotation either.

The key feature of mental quotation is not spatial containment, but rather, a kind of referential or representational constitution. The quoted material partly constitutes the whole mental quotation itself, but whatever functions as the mental quotation "marks" need not literally *surround* what they quote.

---

[12] For the distinction between signs and expressions see Capellen and Lepore (2007, Ch. 12). Roughly, the idea is that different sign systems can be used to articulate expressions of a language, *e.g.,* one can write, speak, type, or sign English expressions. On their view, a quotation may quote either a sign or an expression, but not both at the same time.

[13] This doesn't beg the question against use theories. Rather, the point is that *even if* one thinks that quotations are context-insensitive, that context-insensitivity does not rely on spatial containment.

3.3 Mental Quotation and Context-Insensitivity

Many theorists of linguistic quotation argue that quotation expressions are context-sensitive. They argue that a given quotation expression can refer to different things on different occasions, in different contexts of utterance. For example, Garcia-Carpentero argues that the same quotation can quote distinct items depending upon the context of utterance. On his view, the expression 'gone' can quote: an expression ('gone' is dissyllabic); different types somehow related to the token (*e.g.,* the graphic instance of the uttered quotation or the spoken instance of the inscribe material, *e.g.,* 'gone' sounds nice); different tokens somehow related to the quoted token (What was the part of the title of the movie which, by falling down, caused the killing? 'gone' was); the quoted token itself (At least one of these words is heavier than 'gone', constructed out of large heavy letters (Garcia-Carpentero, 1994, p. 61). He also argues "there are contexts in which the quotations '*Madrid'* and 'Madrid' would have the same content, but there easily conceivable contexts in which they would have different contents" (260). The point is that speaker intentions in a context of utterance are what determine what is quoted on a given occasion.

Context-sensitivity is relevant to mental quotation because, on the view developed in this dissertation, mental quotations (higher-order quotational concepts integrated with appropriate first-order representations) are supposed to determine the phenomenal character of a given mental state; SEEMS <red> determines the red what-it's-likeness of the state. But if the very same representation (SEEMS<red>) can quote different things in different mental contexts of instantiation, that would allow for there being identical mental quotation expressions with different phenomenal character. Phenomenal character

would seem to vary independently of the mental quotation expression, and as a result, the higher-order misrepresentation problem would reemerge. We need to know at what level of abstraction mental quotation is supposed to operate. Whatever the proper understanding of linguistic quotation is, mental quotation cannot be context-sensitive in the same way as linguistic quotation. At this point the distinction between the two possibilities for developing the view should be highlighted.

A quotational concept would have to quote an occurrent experience, the content of which is already determined independently of the mental quoting event. Since non-conscious processes deploy mental quotations automatically, there is not the same potential for variations in usage. Speaker intentions do not determine what gets quoted in the way that they might for linguistic quotations. The content of the first-order representation is fixed prior to its being made available to the higher-order quotation-wielding faculty. The mental quotation latches onto an already existing first-order representational content and generates a concept using that very token first-order state itself, but the content of the first-order state is already determined and fixed. Since the model operates at the level of content, a quotable item does not have the same potential to be "used" by the higher-order faculty in different ways in different contexts of thought, as a linguistic quotation might for a given speaker in a context of utterance. Essentially, (10) is impossible.

> 10) SEEMS<red>, where 'red' is being used to represent the fine-grained appearance properties of green.

One can appreciate how (10) is impossible by recalling the fact that mental quotations need either quote at the level of content or commit to a LOT account of experience. In either case it is crucially the content constituting the experience that gets

quoted. The latter case, in particular, involves quoting a symbol (the LOT expression that constitutes the relevant experience) *and* its content. It is as if the metalinguistic quotation expression 'Fodor' brought the man along with it. This marks a significant distinction between mental quotation and linguistic quotation.

Whatever the proper account of linguistic quotation is, for the account that I am developing here I will assume that quotation is not a merely pragmatic feature. While our method of symbolizing linguistic quotation is conventional, mental quotation as symbolized internally is not conventional in the same way. One would expect it to be hardwired, or at least deeply entrenched, in the structure of cognition. There are various accounts that might explain why it would be, but ultimately it's an empirical question. In Chapter 5 I briefly speculate about one such account. What I am proposing is that mental quotation be understood as a fundamental metacognitive process, which is not determined by higher-level, consciously agreed upon social conventions. In that regard mental quotation is one of the most primitive ways of connecting mind to world.[14]

3.4 The Higher-Orderness, Scope, and Mechanisms of Mental Quotations

Mental quotations are higher-order, or metarepresentational states because they are about other mental states. Since the contents of such states are partly conceptual (mental quotations are quotational concepts), it should be obvious that I am characterizing such states as higher-order *thoughts.* In particular, and as stated above,

[14] Whether or not the linguistic convention arises from the process is a question that I will leave to linguists.

mental quotations are concepts that refer to perceptual states in virtue of embedding them into the same *thought*.

Graphemic quotations have a clearly demarcated scope. For example, if quotation is symbolized with tick marks, the scope of the quotation expression is whatever occurs within the surrounding marks. If quotation is symbolized with italics, the scope of the expression is demarcated by what appears in italicized form. But again, these are conventions. One question that arises is: How is the scope of mental quotation specified?

The scope of mental quotation can be characterized as a function of attention and whatever contents attention determines get globally broadcast. In global broadcasts, information is transmitted to a wide range of nonconscious cognitive systems.[15] Mental quotations are a kind of mental-state term. Thus, one candidate for the system that deploys quotational concepts is the mindreading system, or its evolutionary/developmental precursors. According to this characterization, attention determines which contents are transmitted globally. Those contents are the only ones that can be quoted in a given instance. That is what determines the scope of the quotation. Once the relevant information is transmitted, a concept is formed which uses those very contents to refer to them. That resulting state is the mental quotation.

It is also a working assumption of many theorists that various concepts are appended to globally broadcast percepts as a result of feedback and feed-forward processing between cognitive systems. For example, when I first see the duck rabbit, I may consciously perceive it as a rabbit. Only after the back-and-forth exchange between

[15] The Global Broadcast, or Global Workspace model, was discussed briefly in Chapter 3 (Baars, 1988, 1997; Baars, Ramsøy, and Laureys 2003).

attention, the global workspace, and conceptual systems might I then recognize, and consequently, see it as a duck. We can envision the same kind of processes being true of mental quotation.[16] As concepts get embedded into perceptual representations, so too does the higher-order mental quotation become integrated with a first-order state.

This section presented the main character of the quotational account. In the next section I will illustrate the explanatory power of the model and make explicit it main advantages over existing HOR theories in general.

4. What the Quotational Model Can Do

In Chapter 1 a set of core data were introduced. In Chapter 2 outstanding problems for higher-order theories were introduced. In this section I explain how the model, which is a version of intrinsic higher-order thought theory, addresses the core data introduced in Chapter 1 as well as the higher-order problems discussed in Chapter 2 and 3. Some of the explanatory power of the quotational model is inherited from traditional higher-order accounts, as well as currently existing intrinsic ones. However, as we have seen traditional *extrinsic* higher-order theories run into various problems, and, while existing intrinsic higher-order theories are an advance over traditional extrinsic views,

[16] One might think that inattentional blindness would present a counterexample to the characterization I have just given, *e.g.,* if mental quotations quote, say, the entirety of a visual scene, and being quoted is what makes something consciously experienced, then why is it that in some cases we actually seem to miss certain features? Here I offer a promissory note to show that inattentional blindness does not present a problem for the view. First I need to lay out how mental quotation functions in a theory of phenomenal consciousness. As we will see in Chapter 5, inattentional blindness is actually an illustrative case that further elaborates how the model operates.

they too encounter some of the same problems.[17] The quotational model has the advantage of preserving the virtues of existing views while relinquishing their various shortcomings.

4.1 The Core Data

D1 & D2: What-it's-Likeness & The Distinction Between States that have it and States that Don't

Phenomenally conscious states have a subjective feel, or a mental appearance. They are, as the saying goes, "like something" *for* their subjects. Subjective feel, or what-it's-likeness, is the essential datum that any theory of phenomenal consciousness must explain. Additionally, some mental states exhibit subjective feel, while others do not. In what, exactly, does the difference between these two kinds of state consist? Recall from the discussion in Chapter 1 that an account of subjective feel would presumably tell us how to distinguish between non-conscious and conscious states. With a better grip on what it is that makes conscious states conscious, we could distinguish those states that are conscious from states that lack the relevant consciousness-determining factor (or cluster of factors). However, also recall from Chapter 1 that an account of the conscious/non-conscious distinction would *not* necessarily provide an account of subjective feel. For example, one might be tempted to appeal to independent neural structures, marking off distinctive causal paths leading up to each kind of state, to distinguish a conscious pathway from a non-conscious one (as some do by appeal to the two visual systems

[17] One of these, higher-order misrepresentation, will be discussed in detail in section 3.3.

hypothesis).[18] Doing so provides some account of the distinction between the two kinds of state (they have different causal histories, say, and feed into a distinct set of cognitive systems thereafter), but it would still fail to explain what consciousness itself consists in. It would still fail to explain *why* the conscious path is the *conscious* path.[19] Thus, any theory of consciousness must do more than just distinguish between conscious and nonconscious states. It must distinguish between the two kinds of state *while being able to explain* subjective feel and the other main explananda.[20]

The quotational model can do exactly this. In virtue of having fine-grained/world-representing first-order contents, such states will have qualitative character. For example, a first-order visual state representing a tree will partly determine the worldly properties represented in experience (its apparent color, shape, size, &c.), and representing such features will enable the creature to navigate the world (*e.g.,* it will be able to avoid the tree, rather than bump into it; climb the tree to pick fruit, or what have you). This much can be explained by a first-order theory, but it doesn't tell us what the subjective feel of the experience of the tree for the subject consists in. Following in the wake of existing higher-order theories, the quotational model explains the subjective feel of experience, at least in part, by appeal to the higher-order, experience-representing component of the mental quotation. In virtue of undergoing a representation of that first-order state as a

[18] See Milner and Goodale (1995).

[19] This is not just another "explanatory gap." There is much more than can be said about what makes a state conscious than merely describing its causal history.

[20] As discussed in Chapter 1, this is relevant to a main argument against first-order theories, which may plausibly be held to explain the former, but not the latter. See also Carruthers (2005) and Kriegel (2009).

state the subject is undergoing, the first-order nonconscious state will take on a conscious mental appearance. The experience of perceiving its many leaves fluttering in the wind, their different shades of green in spring, and the darkness or lightness of its bark, will be like something for the subject who undergoes it. Recall, from Chapters 1 and 2, that one intuitive way to draw the distinction between representation and conscious representation is between representations and representations of those representations. But, in contrast to EHOR theories, the higher-order component is neither merely a descriptive thought about the relevant first-order state, nor is it numerically distinct. As discussed in Chapter 2, a numerically distinct state might inform the subject that she is undergoing some or other experience. However, to explain the direct subjective awareness of the state, the experiential quality of undergoing the state, one might plausibly wonder whether something more is required. The problems with EHOR above can be seen as symptoms of the general requirement of extrinsicness. No such problem arises for IHOR theories. The higher-order component is not a description that one is undergoing a FOR, rather the two, when combined, form a unified lens through the world is consciously perceived. Other IHOR theorists have argued similarly. The distinctive contribution of the quotational account is that it also provides a plausible account of the integration between the first-order and higher-order components.

It is important to notice that, in this case, the distinction between nonconscious states and conscious ones is not being drawn merely in terms of two possible causal paths, say, the dorsal and ventral paths. The model (in addition to background assumptions about the relevant neural pathways) tells us that what makes the conscious path the conscious path is that only the percepts following that causal path are available for mental

quotation. This distinguishes the two while also explaining what the subjective feel of that state consists in in virtue of mental quotation, *viz.,* in virtue of the quotational structure, its specific content, and the process of forming a state with that structure, which representations along only that path have.

There are also phenomenological considerations that favor the view. Recall that transparency was originally introduced by first-order theorists, and is sometimes used as an argument against HOR theories. The gist of the transparency thesis is that you *cannot* focus on experiential properties. You see right through your conscious experience only to focus on the things one is conscious of. Phrased this way, which is the typical understanding of transparency, higher-order theorists deny the transparency of experience. For they think that one can in fact be conscious of one's own mental state. But as stated, the transparency thesis is too strong. Transparency is a phenomenological observation, but not everyone agrees on what the phenomenology reveals. For my own part, I think there is an important sense in which experience is modestly transparent, and HOR theorists need not reject transparency altogether. What I find, is not that I *cannot* focus on experiential properties, but more precisely, that when consciously experiencing the world I usually do not. For the bulk of conscious experience is not focused on the experiences themselves, but still, those properties of experience are always in the periphery of conscious experience. Kriegel attempts to capture this idea with the notion of "peripheral self-awareness." I think the idea has merit; however, I don't think he fully captures it. A brief discussion of what I have in mind will highlight one significant difference between the COII account (discussed in Chapter 3) and the quotational account.

On Kriegel's view a higher-order representation is integrated with a first-order representation. There are in fact two distinct representations that are part of the same state: a first-order state M and a higher-order state M*, which represents M. But notice, while the first-order state and the higher-order state are parts of the numerically same state, they are actually two different representations. Kriegel explicitly says the two states start as "initially separate states" (2005, 46). On the quotational account, there is no *separate* higher-order state that forms part of the complex. The higher-order state represents the first-order state by latching onto that very state. When considered severally, the higher-order component is not a complete representation; it is merely a concept frame. In contrast, when considered severally, the higher-order component of one of Kriegel's COII states is a complete HOR. It would be a typical extrinsic HOR were it not for being "integrated" with a separate first-order state. Given the robustness of the higher-order component of a COII state, it is difficult to see how transparency could be captured. One might wonder why the self awareness is only peripheral, such that transparency should be rejected altogether. In contrast, the quotational account captures the modest transparency of conscious experience in that the HOR really is, in a sense "transparent." The higher-order quotational state simply latches onto an already existing first-order experience. The higher-order content is, in a sense "seen through." However, since it both represents the world and the FOR as an experience the subject is undergoing, the peripheral self-awareness remains in the background. It is in that regard that the state has both a "for-me-ness" and a blueness, say.

D3: Intimacy

As discussed in Chapter 1, there is a family of notions traveling under the heading "intimacy," the members of which should be kept largely distinct. Typically, though, the apparent intimacy of conscious states is thought to be a problem for representational theories and for higher-order representational theories in particular.

The quotational model is particularly well-suited to deal with intimacy claims, accepting that we have a genuine intimacy (in any of the above mentioned senses) with our own conscious states. On the view of quotation being put to work in this discussion, a quotation expression cannot quote anything other than what it does in fact quote. On this view, mental quotations employ a kind of constituting-representation, *i.e.,* they are partly constituted by the tokens of the experiences they represent. Barring strict identity, it is difficult to see how a relationship could be any more intimate than that. Consequently, the quotational model can acknowledge and explain why conscious states exhibit intimacy.

First, according to the quotational model, both the object of conscious awareness and its targeting state arise in direct simultaneity as the one conscious state. Thus, the model accommodates the temporal reading of "intimacy." Second, the model also explains why self-discoveries on the basis of a therapist's testimony do not on their own result in a phenomenally conscious state. The results of such discoveries tend to be descriptive; they characterize oneself as feeling a certain way, but they do not tend to involve an active token of a given feeling (as the object of the awareness). This is not to say that such discoveries cannot lead to a phenomenally conscious state. They certainly could if they lead to a mental quotation of the active state that is discovered. But it's unclear why that *must* be the case, and in fact, it seems like it usually is not. As for the

second notion of "noninferential," relatedly, the answer is that one cannot know what it's like to undergo a particular conscious experience on the basis of inference alone, *if* the inferential process generates a typical (non-quotational) conclusion, because (as is suggested by the situation of Mary) the subject lacks the actual fine-grained first-order content which constitutes the (non-conscious) experience. But again, *if* it were possible to generate the relevant quotational state as a result of inference, then one could know. Overall, there is nothing about inference per se that rules out its product being phenomenally conscious. What seems more likely is that our system happens to work that way, but there is nothing in principle that rules out an alternate inferential system, the outputs of which are mental quotations (phenomenally conscious states).

4.2 Higher-Order Problems

4.21 Rocks & Other Minds

Recall that existing intrinsic views can handle the rock problem better than extrinsic theories. I am not going to review the ways in which extrinsic views can deal with the problem here. I discussed that in Chapter 2. Here I will remind the reader that the discussion in chapter two illustrated that the problem is not as threatening to extrinsic views as it is sometimes assumed to be. There are quite plausible things the extrinsic theorist can say. Existing intrinsic views, on the other hand, are even better suited to deal with the rock and other minds objections. For intrinsic views the problem does not even arise. The integration that such views require rules out the objections from the start. The problem that I raised for existing extrinsic views in Chapter 3, was that none provides an entirely satisfying account of the integration between the requisite components of a

conscious state. In that regard, the quotational model can handle the objection better than existing views.

On the quotational model a complex consisting of an appropriate first-order state and a higher-order quotational component is what makes a mental state conscious. The quotational model predicts that rocks and inanimate objects cannot be conscious for the simple reason that rocks and inanimate objects are not the relevant kind of complex state (consisting of integrated first-order and higher-order components). While a mental state of another subject is the right kind of thing to be a constituent of the requisite complex, it cannot be the object of a complex in another person. Notice the "cannot" here is not metaphysical. It expresses natural necessity. If we could physically get an appropriate FOR from one person to be part of a complex quotational state of another person, the quotational model predicts that such a state would in fact be conscious. However, this is not as bizarre as it might seem at first glance. It is about as bizarre as the idea of a heart in one subject also sustaining the circulation (and life) of a second one, *if it were physically possible.* Naturalists should not have a problem with either possibility.

4.22 The Too-Easy Problem

In Chapter 2 I argued that EHOR theories have a ready response to the too-easy problem. This is one of the explanatory features that the quotational model inherits from existing EHOR views. In fact, the quotational model may be even better suited to handle the too-easy problem. One might argue that quotation is even easier than ordinary higher-order thought, but that would be to miss the point of *mental* quotation, which, as I have illustrated, is different than linguistic quotation in significant ways. I will return to this issue in Chapter 5.

4.23 The Problem of Higher-Order Misrepresentation

Several authors have argued that higher-order theory, in particular, (by which they mean *extrinsic* higher-order theory) encounters problems handling intimacy, because of the "distance" between the first-order and higher-order states (Goldman, 1993; Natsoulas 1993; Kriegel, 2006, 2009). The one particular problem that is sometimes framed as an intimacy problem is the puzzle of higher-order misrepresentation. As discussed in Chapter 2, the problem was first introduced by Karen Neander (1998) and Alex Byrne (1997). Since then it has reemerged in various, but closely related, forms.[21]

In Chapter 2, I examined some of the ways the traditional extrinsic higher-order theorist can respond, and concluded that none is very plausible. The two most promising ways are what I called the "conjunction" view, according to which matching higher-order and first-order components are set as a constraint in a non-ad hoc manner, and what I called the "intentional inexistents view," according to which a higher-order representation itself, in the absence of a corresponding first-order state, is sufficient for the subject to undergo a phenomenally conscious state. There I argued that both of these options fail for the same reason: they fail to account for the fineness of grain of conscious experience.

For those who remain sympathetic to the higher-order framework, the problem of higher-order misrepresentation motivates the self-representational theory of consciousness, or what I call "intrinsic" higher-order theory. Intrinsic higher-order theorists would seem to have an easier time avoiding intimacy problems, because they tighten the distance between the first and higher-order components, maintaining that both

[21] See also Levine, (2000); Kriegel (2009); Block (2011).

components are parts of the same state. However, as discussed in Chapter 3, existing intrinsic views encounter difficulties addressing the very problem they set out to solve. They either fail to rule out higher-order misrepresentation altogether or they introduce arcane commitments to do so, at the cost of complicating the view well beyond necessity.

Intrinsic accounts are offered, at least in part, as solutions to the higher-order misrepresentation problem, but none adequately handles the problem. The quotational model is an alternate version of intrinsic higher-order theory, and as such, it holds that in a conscious state, the first-order state is a constituent of its higher-order (in this case quoting) component. In contrast to competing intrinsic views, the model provides an account of the integration of first-order and higher-order components that does in fact rule out higher-order misrepresentation. To review, a higher-order mental quotation represents its corresponding first-order state in virtue of presenting that very state as a state its subject is undergoing. To phrase it somewhat more dramatically, the higher-order quoting component "latches onto" an actual activation of the very first-order state that it represents. This makes subjective feel, or phenomenal character, strictly parasitic upon qualitative character. Since phenomenal character is partly comprised of qualitative character, then if the relevant corresponding qualitative character is absent, there cannot be any phenomenal character.

One reason to think that phenomenal character is parasitic upon qualitative character in this way has to do with the fineness-of-grain of conscious perception. It is plausible to think that qualitative character does not arise independently of an appropriate first-order representation, *e.g.,* one with partly non-conceptual (finely grained) contents with mind-to-world direction of fit. Moreover, it is plausible to think that such a state

typically does not arise independently of something like a prototypical sensory system. Thus, if there is no first-order sensory state (or something very much like one) to be mentally quoted, there simply cannot be any phenomenal character.[22]

If the above line of reasoning is accurate, higher-order misrepresentation is simply impossible on the quotational account. Higher-order representation by means of mental quotation just doesn't work that way. But, as we have seen, this is not a mere stipulation of the structure of a conscious state. It actually *explains* the distinctive feel and intimacy of conscious experience, among other distinctive features of consciousness, while simultaneously ruling out higher-order misrepresentation.

5. Conclusion

This chapter presented the central features of the quotational model of consciousness and examined the ways in which it handles the initially puzzling data introduced in Chapter 1 and the specifically higher-order hurdles introduced in Chapters 2 and 3. In the next chapter I will work through some challenges for the quotational model as well as some future directions for the view.

[22] One caveat is that a targetless higher-order *perception* might well be able to manufacture finely grained first-order content on its own. If so, a targetless higher-order perception might well be like something, however, there is no reason at all to think that there is such a higher-order faculty of inner sense.

Chapter 5: Conclusion: Objections, Replies, and Future Directions

This chapter addresses objections to the quotational account and suggests some future directions for the view.

1. Objections and Replies

1.1 Just What is This Mechanism of Mental Quotation?

One might wonder how the mental quotation "mechanism" works–especially given that ordinary quotation exploits lots of pragmatic conventions. What is the computational procedure that will ensure the crucial semantics required of mental quotation?

What I am calling mental quotation is *importantly different* from quotations used in both natural language and artificial languages. In natural language, for example, "Fodor" is a quotation expression, but Fodor the man isn't contained between the quotation marks. Rather, there is a linguistic symbol, i.e. 'Fodor'. In the previous chapter I outlined two ways of spelling out the quotational view. According to one, mental quotations quote symbols in a language of thought. Among those symbols are the very sensational symbols themselves. On the other hand, one might describe mental quotation operating at a different level of abstraction, as a concept that quotes experiences or contents themselves—*not symbols*.

According to the first characterization, the content of the first-order symbols would also have to be quoted as part of the entire LOT expression. This kind of quotation has not been realized in natural language or computer science. I have already given an example from natural language (the "Fodor" example above). Consider the way quotation works in an artificial language such as Scheme. There are various reasons why one would want to use the quotation operation in artificial languages. One is to distinguish between, say, a valid procedure and a list. For example, suppose you want to distinguish the procedure application (+ 1 2) from a mere list. To treat it as a list we use the "quote" procedure: `quote`.

`(quote (+ 1 2)` ⇒ `(+ 1 2))`

`quote` tells Scheme to treat the relevant list itself as data rather than as a procedure application.

Here's another textbook example of artificial quotation. Suppose you want to use a word itself, say 'square', rather than its value. If you enter `square` , Scheme thinks you want to run a procedure (the square procedure). To use *the word* 'square' we need to tell Scheme we want to use the word itself, not the value of the word. Again, we do this `quote`.

`(quote square)` ⇒ SQUARE

The point here is that quotation in Scheme is a clearly defined process that *ignores the value of the word and uses the word itself.* This exactly not what I am claiming *mental* quotation is like.

Above I argued that the quotational model can be developed in purely LOT terms, so long as one also agrees that sensations themselves can be characterized in LOT terms

too.[1] As it stands, the above procedure would not capture mental quotation, which would, in LOT terms, employ both a "word" *and its value.*[2] The (merely) analogous "mental" quotation procedure would be something like the following.

```
(Quote (square) and square)
```

Notice, I am using the Scheme example here merely for heuristic purposes. I am not claiming that a scheme quotation that quotes a symbol and its value would thereby be conscious![3] Nor that such a procedure actually exists.

There might be some reason why you would want to employ the above hypothetical procedure with a word in Scheme, but it's not clear to me (and I don't think it's a recognized function in the language). I leave that to programmers, who are charged with finding uses for various procedures of the languages within which they work. Meanwhile, there is in fact a way to specify the operation of mental quotation *at the intentional level*. As stated above, instead of endorsing a LOT account of sensation, one might choose to develop the quotational account at the level of content, not symbols. On such a view, mental quotations are concepts that take experiences (fine-grained contents) as their objects. As I say at pp. 120-121, there is an important contrast between mental quotation and linguistic quotation. This is it. On either construal of the quotational account, *mental* quotation is importantly different than linguistic quotation. It either

---

[1] Typically the language of thought hypothesis is silent perceptual states however, the LOT view of perception/sensation does have its proponents proponents. See, *e.g.,* (Rey, 1992, 1993) and Vineuza (2000).

[2] At one point, something like this was Davidson's view of linguistic quotation, *i.e.,* linguistic quotations both use and mention.

[3] I have already addressed this issue in Chapter 2 and in Chapter 4. However, it will be revisited once again in section 2.

quotes a symbol and its value (if one endorses the LOT account of sensation), or it quotes at the level of content in the first place.

One might object that I still have not specified the computational procedure for the mechanism of mental quotation. But the quotational view is building upon existing higher order theories, none of which proposes a computational procedure. Higher-order theories are presented at the level of intentionality, not computation. Of course, we ultimately want an explanation of how intentional procedures are realized by computational ones. For the purposes of spelling out an alternative IHOR theory, it's not my charge to provide the actual details of that computational procedure. I do think that an intentional model can at least *tell us where to begin to look* to develop a computational model. To insist on a computational procedure from the start gets the order of explanation backwards: we need to know something about the structure of the thing before we start describing it mathematically. One way to construe higher order theories is: they're trying to provide the intentional model on which a computational model might possibly be based *at some point in the future*.

1.2 Quotation: is Still "Too Easy"

One might claim that most computational proposals in this area essentially involve states and processes that could be easily realized on a laptop. In Chapter 2, I discussed the "too-easy" problem as a problem for EHOR theories. One might wonder whether the quotation proposal is especially vulnerable to the too easy problem, since mere quotation is especially trivial and prevalent in existing computing systems. Moreover, so the objection goes, one cannot exclude this "too easy" objection by

restricting the quotational account to perceptual states, because most existing laptops these days have microphones and video recorders built in. Let the outputs of these devices get processed by some Baysean computations that are the favored model of perceptual processes these days –and let us suppose we also supply come computations that allow the inputs to be quoted, et voilà!?? It would appear that the quotational account of consciousness would be committed to the laptop as having auditory and visual experiences!

In response, let me remind the reader that the too easy problem was dealt with in Chapter 2. There, I argued on behalf of EHOR theories that they already have the resources to address the problem. The quotational version of IHOR is no worse. More specifically, the "it's still too easy" objection weaves together two different points that should be kept distinct. First, it charges the quotational account of attempting to exclude the too easy objection by restricting the account to perceptual states. However, the quotation account does no such thing. The quotational model is restricted to perceptual (actually sensory) states simply because such states are the paradigmatic conscious states. It may be that attitudinal states can be conscious too (the kind of state that is put to work in the Freudian component of the too-easy objection), but, since there's disagreement about this, let's set those more controversial states aside and focus on the states *everyone agrees* may be conscious (if any are conscious at all). In Chapter 5 I propose, as a future project, a discussion of whether the quotational account generalizes beyond perceptual states to attitudinal states. But I see no reason why I must provide that as part of the *present account.* The restriction to perceptual states does allow the quotational model to set aside the Freudian component of the too-easy objection, but that

is a distinct worry with a different answer. In that regard, the model is no worse than existing HOR theories. The Freudian objection simply doesn't apply to the account I'm offering. The charge that quotational states are prevalent in ordinary computing systems, on the other hand, does. But, to reiterate, these two components of the objection should be kept distinct.

To respond to the ordinary computing systems component one must appreciate two points. First, while ordinary computing systems, say laptops, may realize linguistic quotation states, they don't realize *mental* quotation states (neither the LOT formulation nor the intentional level formulation). But notice, even if one did, that alone wouldn't count against the quotational account. The system would fall short of a significant constraint. To stick with the example mentioned above, the laptop/camera/mic device doesn't meet the criteria of a prototypical sensory modality. I outlined these on p. 50 and p54, and originally in Picciuto and Carruthers (2011), but similar constraints can be found all over the place in the recent senses literature (see for example, Macpherson, 2011).

Recall that a prototypical sensory modality will: (1) be sensitive to some range of physical energy or set of physical properties, (2) include a detector mechanism that transduces that energy or those properties into informational signals sent to the central nervous system where (3) they are used to guide the intentional behavior of the organism. In addition, a prototypical sense will (4) have as its evolutionary function the detection and representation of the physical energy or properties in question, and (5) will issue in nonconceptual representations with mind-to-world direction of fit. Again, it's not clear how much 4 really matters, but this is the outline of a *prototype*. It seems clear that the

laptop doesn't fit the bill. It barely meets (2), and there is no reason to think that such a machine meets (4). Additionally, the system clearly lacks a concept of 'appears' (or at least the proponent of the objection has given no reason to think that it has one), but that is one of the essential features of a *mental* quotation. Thus, the objection betrays a misunderstanding of the core thesis of mental quotation.

Moreover, I am willing to concede that *some* artificial systems could be conscious, but I do not think I am committed to the claim that even a modified laptop (at least not one modified in the way suggested above) is conscious. This is why in Ch.2 I skip to considering a sentient robot. It might be an interesting future project to carefully examine the range from "least sophisticated system" to "most sophisticated system," hypothesizing about the different degrees of consciousness that might be present along the spectrum. But I don't see why I need to do that here. Until the proponent of the objection provides convincing reasons to think that modified laptops realize fine-grained perceptual states, this component of the too-easy object does not present a problem for the quotational account. Merely asserting they do is insufficient. It begs the question against views that would want ascribe consciousness to minimal systems (not that mine does that…).

Furthermore, the quotational account is intended as a reductive explanation of phenomenal consciousness. According to any such view, consciousness is nothing mysterious. But why then should we think that it is so mysterious that an appropriately modified artificial system be phenomenally conscious? Because it runs counter to our folk-intuitions? That certainly isn't enough. Again, these intuitions might lean toward the biological, but it's unclear just how much such intuitions are worth.

Another response to which I am open acknowledges that consciousness comes in degrees. The transitivity principle (Ch.2) already says that that conscious states are states we're conscious of being in. With the transitivity principle in place, phenomenal consciousness can be construed as a function, not *only* of being higher-order represented, but also of there being an appropriate first-order representation. But note, part of what will determine the phenomenal character of the complex mental quotation is the content of the very first-order representation it targets (phenomenal character is parasitic on fine-grained qualitative character). Whatever kind of state the laptop/camera system realizes, even if it can in fact be shown that it's a fine-grained perceptual state, and that the system undergoes the relevant kind of quotational state, its content isn't very robust at all. This is partly in virtue of its lacking myriad concepts, but also in virtue of lacking a first-order state with much *content* at all in the first place. On the quotational model if there is no fine-grained content to be conscious of, then there is no conscious state. Similarly, if the first-order state possesses minimal fine-grained content, such that there is hardly anything to be conscious of, then the phenomenal character will be dim, on my view. So even if I do concede that such a system would be phenomenally conscious (which I have already argued I do not have to do) that system would be phenomenally conscious *to a minimal degree*. One can appreciate this point by considering the phenomenology of peripheral vision, or of say, states of partial creature consciousness that we actually undergo. It is important to appreciate that this is not merely bullet biting of the sort offered by Lycan, Prinz, and Papineau. It's what the theory predicts and actually does explanatory work.

1.3 The Unconscionable Slide from Symbol to "Experience"

One might think I am moving from a more basic representational state (my use of 'mental representation') to a so-called "experience," and that in doing so I am sneaking something that remains unexplained along the way.

I have two replies to this objection. First, I am dealing with a deflated notion of experience. I reject Block's claim that phenomenal consciousness just is experience. As many others do, I acknowledge both conscious experience and non-conscious experience (I discussed this in Chapter 1). All I take a non-conscious perceptual experience to be is a perceptual representation (a mixed conceptual/non-conceptual representation with mind-to-world direction of fit, as outlined on pp. 50 & 54). These are typically the outputs of a prototypical sensory system, but in principle they wouldn't have to be.

Second, this objection fails at a much deeper level. For it seems to mischaracterize the quotational model as sliding from quotation of a mental representation, or symbol (because that is what linguistic quotation is, and on the view being offered I am just importing the notion of quotation from the domain of language to the domain of the mind without change) to quotation of a perceptual state (which may be construed as a symbol, with content, and causal/functional role). However, I am not merely importing linguistic quotation (or some process that is isomorphic to it) from language to mind, without change. As I have stated, mental quotation is importantly similar to linguistic quotation, but it is also significantly different. I simply reject the premise of the argument.

1.4 "I Am Not Now Having That Experience"

In Chapter 4 I discussed the origins of the idea of mental quotation. To review, Papineau introduced the idea as an account of phenomenal concepts, but then gave up on the idea of mental quotation. His main reason for abandoning the view seems to center on the following objection.[4] It seems like one can truthfully think the thought 'I am not now having that experience, nor recreating it in my imagination' (Papineau, 2006, 2). For example, if one is not actually undergoing, say, the experience she had last Tuesday, then one should be able to truthfully think 'I am not now having *that* experience (nor recreating it in my imagination)'. And, so the objection goes, if in thinking 'I am not now having that experience (nor recreating it in my imagination)'one must deploy a phenomenal concept, then it seems that the quotational view makes it impossible to truthfully think that thought, because doing so would involve a tokening of the very experience that one is denying presently to undergo. Moreover, it seems like we can truthfully think that thought, so it would count against the view if it entailed that such a thought is actually impossible.

This objection raises a problem only if it is assumed that in thinking the above thought one *must* deploy a phenomenal concept. But why think that? There are any number of ways to refer to a prior experience between which the thought expressed by the surface grammar 'I am not now having *that* experience (nor recreating it in my imagination)' is ambiguous. It is not the case that one must deploy a phenomenal concept to truthfully think that thought. In some instances of a thought with surface grammar 'I am not now having *that* experience (nor recreating it in my imagination)', the expression

[4] According to Papineau (2006) the objection was raised by Tim Crane in conversation.

'that' may function as an ordinary demonstrative, and not necessarily as a phenomenal concept. When 'that' functions as an ordinary demonstrative in the above thought, it merely points to the experience in question, but it does not require an activation of that very experience (neither the very same token, nor a token of the same type). Moreover, even if the ordinary demonstrative tokening of 'that' does lead one subsequently to recreate imaginatively the experience, it need not do so.

The main point is that a thought with the above surface structure can have both true and false tokenings, depending on the underlying structure in the relevant case. A quotational tokening would be false. An ordinary tokening would be true. The main worry was that the quotational account makes it impossible to think (truthfully) the seemingly true thought 'I am not now having *that* experience (nor recreating it in my imagination)', but mental quotation does no such thing. While it is true that quotational tokenings will in fact be false, that is not the only way to think that sort of thought. There is no reason why we should think that we must always deploy phenomenal concepts when thinking about our own experiences. Thus, since the counter-exemplary thought does not require the deployment of a phenomenal concept, the objection is not decisive against the quotational account of phenomenal concepts.

1.5 Change Blindness

In the previous chapter I mentioned that cases of inattentional blindness could, at first glance, be seen as counterexamples for the quotational account. In this section I will argue that they do not. I will also use inattentional blindness to further develop how mental quotational might function.

Cases of change blindness seem to involve the whole of a visual scene being quoted, but yet there are features that go unnoticed, or that are not part of *conscious* experience.[5] How can this be, if a representation of the entire scene is mentally quoted? The answer settled on bears directly on the debate between those theorists who think that experience is "rich" and those who think that experience is "sparse." My gloss on the issue is that while there might be stages of visual experience that are representationally rich, conscious visual experience can still be phenomenally sparse.

In the change blindness paradigm a series of images is presented in rapid succession. These can be drawings, paintings, photographs, or videos. The phenomenon has even been evoked in scenarios with live players (gorilla suit) and in uncontrolled form in, *e.g.,* the illusionist Derren Brown's "person swap" sketch, though the latter is anecdotal.

In experimental scenarios using the flicker technique, the first visual image is presented followed by a blank screen that is followed by an image, which might or might not contain a difference from the first. The difference can be subtle or quite large. The series is then repeated. Subjects consistently fail to notice the difference even after viewing several repeated cycles of the series. For example, in Figure 1 the shadow from the helicopter is missing from one of the pictures. In Figure 2 the railing is slightly higher in one scene.[6] If one can notice the change at all (in some scenarios it is more difficult

---

[5] Whether or not they are constituents of *conscious* experience will depend in part on whether one thinks that conscious experience can be inaccessible. For example, Block would dispute this characterization of change blindness. Block's own interpretation will be discussed below

[6] The Change Detection Database contains several additional examples evoking the change blindness phenomenon http://viscog.beckman.illinois.edu/change/info.shtml.

and requires more time to discern the difference than in others), it typically takes multiple viewings for subjects to consciously notice the difference. There are at least two ways to interpret these results. One can claim that when subjects miss the changed features that these features are not perceived at all (Dehaene *et al.* 2006; Noë, 2004, O'Regan and Noë, 2001). This is the view that describes such cases as cases of inattentional *blindness*. On the other hand, one can claim that when subjects miss the changed features that these features are consciously perceived but, nevertheless, that subjects fail to notice them (Block, 2001; Dretkse, 2004). The latter view describes the situation as involving inattentional *inaccessibility*. According to the inattentional inaccessibility interpretation, the unnoticed features are present as features of the phenomenally conscious experience, but they are inaccessible.

The quotational model explains change blindness in the following way. Initially, the relevant feature that goes unnoticed is not retained by attention in working memory from one presentation to the next. If so, then the subject cannot compare it with the corresponding location, or aspect, of the current experience. Essentially, the changing feature is not attended. Because it is not attended, it is not globally broadcast. And if it is not globally broadcast, it is not quoted.

However, as part of the back and forth processing that appends concepts to visual percepts after which they are globally broadcast, given the right cues or enough time attention might focus on the initially absent features thereby feeding them back into the global broadcast. Since the global broadcast will now include that feature in the original state, once passed on to the conceptual mechanism charged with deploying quotational concepts, it will then be conscious.

1.6 Are Mental Quotations Phenomenal Concepts?

One question that naturally arises is whether the quotational concepts at work in the quotational theory of consciousness really are quotational *phenomenal* concepts. In other words, one might accept the quotational account of consciousness but deny that the quotational concepts that partly constitute conscious states are phenomenal concepts. This issue is largely terminological, for there is only one main substantive reason why the quotational concepts employed by the quotational theory ought not to be considered 'phenomenal', and that is if by employing such concepts in a theory of the nature of consciousness, the phenomenal concept strategy itself is undermined. As stated in the opening passages, the phenomenal concept strategy is intended to defend physicalism against various anti-physicalist arguments, and the core insight of the strategy is thought to be that it distances the nature of consciousness itself from the concepts we use to think about conscious states.[7] According to the view that I have introduced, the conceptual distancing between the explanation of consciousness itself and the apparent mystery of consciousness seems to have been shortened: quotational concepts are partly constitutive of phenomenally conscious states themselves. Thus, one might argue that I have reintroduced the mystery that the strategy intended to explain away. It needs to be shown, then, that the quotational account of consciousness retains whatever explanatory power the phenomenal concept strategy has, else we would have to give an alternate defense of physicalism.

---

7 See Balog (2009).

Fortunately (if the strategy has promise at all), no such alternate defense of physicalism is required; nothing about the quotational theory of consciousness itself undermines whatever explanatory power the strategy has. Consider one example: the quotational theory of consciousness retains a physicalist explanation of Jackson's Mary scenario. Assuming that Mary has progressed through a normal process of maturation (with the exception of living in a solely black and white world), even before she exits her room she would be able to think (and would have undergone) other quotational thoughts. That is, she would have undergone other phenomenally conscious states, *e.g.*, states representing black and states representing white. What she gains when she steps outside the room is not the quotational structure itself. As hypothesized above, that would already have developed or been acquired. Rather, what Mary gains is a specific experience of red that can immediately be integrated into that structure, and thus, rendered conscious. And once Mary has undergone the *conscious* experience of red, she can use that state to think about her conscious experience, whereby she might think something like 'Ahhhh. That is what it's like to (consciously) experience red.' Regarding Mary's situation, nothing about the original alleged explanatory power of the phenomenal concept strategy has been lost. Furthermore, there is no reason to think the explanatory power of the strategy would be lost regarding explanatory gap arguments or zombie/invert intuitions. Thus, if the phenomenal concept strategy is successful, it remains so even under my proposed revision.[8]

8 Again, the issue here is not whether the strategy actually has any promise. The issue is whether or not my revision undermines whatever explanatory power the strategy might be thought to have.

Moreover, it is not merely that the phenomenal concept strategy itself has nothing to lose from being held in conjunction with the quotational account of consciousness; it has something to gain as well. First, unifying an account of the way we think about conscious states with an account of what constitutes state consciousness at all renders the phenomenal concept strategy less ad-hoc. On the quotational account, it is not just that we have unique concepts that we use only to think about states that are already conscious by some independent process (why would we need such unique concepts just for that?), but more importantly, we have those unique concepts *and they partly constitute conscious states* in the first place. The uniqueness of consciousness instantiated by conscious states themselves calls out for unique concepts in a way that merely thinking about such states does not. Also, it offers an explanation for exactly why it is that conscious states *seem* mysterious. Similarly, they seem mysterious not simply because we use unique concepts to think about them, but rather because the unique concepts we use to think about them partly constitute what makes those states conscious in the first place. That is, conscious states themselves are at least *prima facie* mysterious, however, on the quotational view that mysteriousness is explained rather than merely explained away.

1.6 Why Would There Be Mental Quotation?

I have argued that mental quotation plays a crucial role in a theory of phenomenal consciousness. Although mental quotation is a reasonable theoretical posit because it can explain phenomenal consciousness, it would be better if we had independent reasons to

believe in it. Moreover, one might argue that it is a prima facie problem for the view why mental quotation should occur?

One possible way to provide independent support for the quotational model is to situate mental quotation within an evolutionary framework. Admittedly, that requires engaging in speculative evolutionary psychology, however, the speculation has empirical foundations in the data that has been gathered over the past thirty years or so by experimental psychologists and other researchers studying animal minds and behavior. More recently, Lurz (2011) has put forth a similar argument focused on a debate regarding mindreading in non-human animals.

It is plausible to think that the ability to attribute appearance states to others provides humans and other creatures capable of making such attributions with an adaptive advantage. Attributing appearance states would allow an organism to predict other agent's behaviors in situations in which a behavior-reading competitor could not. As Lurz (2011) points out, this would give appearance attributors an advantage particularly in illusory settings, such as when an insect appears to be a leaf. If it appears to A that there is a leaf in her immediate vicinity and A knows that the appearance differs from how things really are, *i.e.,* the leaf is really an insect, then A might be able to use this information to her advantage. In particular, she could use it to avoid a potentially harmful encounter with another agent. She would be able to do that by grasping the appearance/reality distinction and by attributing an appearance state to her competitor. A might think "B sees the insect as a leaf" and conclude that B will leave the setting. A could then snatch the insect without interacting with B. On the other hand A might think "B sees the insect as an insect" and conclude that it is better for her A to move on instead

of engaging with the competitor. In either scenario the capacity to grasp the is/seems distinction and attribute appearance states to other agents can be advantageous.

Compare the appearance attributor to a behavior-reader C, an agent incapable of grasping the is/seems distinction, and thus, incapable of making appearance-state attributions, to A. When C is confronted with the leaf insect and a dominant conspecific B, C could only conclude that B sees an edible insect and this would not help the subordinate predict that the dominant conspecific might not eat the insect (Lurz, 2011). She would do this based solely on observations of her competitors behavior and position (whether the competitor has a direct line of gaze or the proper orientation). In such a situation, it is likely that C would assume that B will go for the insect and leave the scene. But suppose B, the dominant conspecific, actually sees the insect as a leaf. In that case, B might well leave the insect untouched. C will have missed an opportunity to easily obtain food. If C had the ability to grasp the appearance/reality distinction and to make use of that information, C would have had little to do to obtain a meal safely (remain hidden in a bush, say, or maintain his subordinate posture). This, Lurz argues, gives the appearance mindreading creature an obvious advantage over the behavior reader. In order to perform such acts of mindreading, the creature must be able to understand the is/seems distinction and use that information to form predictions.

But there is also an advantage to be had to the creature who can self-attribute appearance states to oneself. Consider the creature that can discriminate potential food or prey that is nontoxic from an imposter which "mimics" the properties of a toxic counterpart (or vice versa). For example, the viceroy butterfly (*Limeniti archippus*) mimics the appearance of the toxic monarch (*Danaus plexippus)*. Some plants seem to

engage in different kinds of mimicry as well. There are orchid species that look like female bees to attract pollinators (Launchbaugh and Provenza, 1993). Other plants resemble stones to avoid being eaten (Wiens, 1978, Barrett, 1987). It seems that some species have evolved to instantiate an appearance reality distinction. The predator that is capable of grasping the appearance reality distinction would have an adaptive advantage. Of course there are other ways that creatures have evolved to deal with imposter prey, *e.g.,* mammalian herbivores can associate post-ingestive consequences with flavor, a cue upon which, they might form a generalization that is reinforced by their feeding environment (Garcia, 1989). Grasping the appearance reality diction is surely not required for survival. However, the creature who is capable of drawing the distinct would have an advantage, particularly in situations wherein it is useful to hypothesize about a competitor's own mental states.

Notice that understanding the is/seems distinction requires more than merely being able to "see as." To see as one must be able to bring different concepts to bear on the same thing, but one could do this non-consciously and without any grasp of the appearance/reality distinction. In fact, these discriminations might even map onto the way things appear and the way things are and that still would not make it the case that the creature grasps the is/seems distinction. To grasp the is/seems distinction, one must have a concept of appearance.

Lurz (2011) gives one possible account of the psychological mechanisms that would underwrite the appearance reality distinction. First recall that it is commonplace to hold that perceptions are a primary source of belief. Perceptions are designed to fix beliefs quickly and powerfully. In an illusory setting my visual system might tell me

there is a leaf in directly in front of me, while my background beliefs tell me that there is an insect directly in front of me. In other words, I can see the insect both as a leaf and as an insect. How might my system resolve this tension? One way considered by Lurz is that in cases of conflict, perception wins by default. However, this is a peculiar result in that the system would resolve to have me (sincerely, actually) believe that there is a leaf in front of me when I know that there is not. Also, notice that many of the background beliefs that constitute my "reasons" for knowing that there is an insect in front of me will themselves be perceptual beliefs. If so, then the notion of perception winning by default must be clarified. It is not that perceptual beliefs trump by default, for there are several other perceptual beliefs at work in one's network of beliefs. More precisely, it would have to be the most recent (active, or occurrent) perception and the beliefs it generates that win by default. It's difficult to see how this could be of much use to the organism in drawing an is/seems distinction, since much will depend on what the creature has perceived most recently, or is currently perceiving, rather than on what the creature has reason to believe, or fits best with the creature's background beliefs and perceptions. Lurz proposes that nature might have designed the perceptual system to generate "qualified" beliefs by default. These are beliefs qualified by "seems" or "appears."[9] Thus instead of the perceptual system generating an unqualified belief that would determine the creature's actions, the system would generate a qualified "seems x" belief by default. The qualified belief would then be compared to background beliefs. If there were no

[9] Apparently there is some reason to think that nature has rigged up certain creatures with the capacity to affix other kinds of belief qualifiers, including temporal, spatial, quantitative, exclusory, and probabilistic qualifiers (Gallistel, 1990; Boysen, 1997; Beran, 2001; Correia *et al*, 2007; Aust *et al*, 2008; Hoffman *et al*, 2009; Beran, 2010; Grodzinski and Clayton, 2010).

conflict, then an unqualified belief would be generated and it would play the characteristic belief role in the subject's system. If there were a conflict, say, in an illusory setting, then the system would not generate an unqualified belief at that time. The subject would only have a qualified belief about how things appear and this would not necessarily motivate the subject to act.

Admittedly, this is only one of at least two stories one can envision. For example, one could allow the outputs of perception to give rise to beliefs by default, rather than appearance states, except where there are existing beliefs that conflict with the new ones. When a conflict is detected, the belief is suspended until it is resolved. This might well be more parsimonious. For the quotational account must also postulate a conflict-detector. For the quotational account case it will be the absence of any detected conflict that allows the "seems" operator to be removed.

It is difficult to say which one is a better explanation. But here too I don't need to decide between the two presently. It is enough to illustrate the version I have described is at least possible (and not implausible).

I have been discussing one view of why certain organisms might have acquired an understanding of the is/seems distinction. Lurz goes on to develop an account of how an organism would come to use that understanding in making appearance attributions. My purpose here is not to resolve any debates about animal mindreading, but rather to illustrate that there is a plausible story about the adaptive advantage of understanding the is/seems distinction and propose plausible evolutionary origins for mental quotation. Notice that while Lurz does not explicitly call perceptually qualified beliefs meta-representational, they are meta-representational. They have first-order, world-directed

perceptual contents, but also, insofar as they have qualifiers directed at first-order perceptions themselves, they have paradigmatic higher-order, experience-directed components. This amounts to the claim that *understanding* the is/seems distinction requires, among other things, meta-representational concepts of appearance (or experience). One needs to know that things seem (are seeming to me) one way but could in fact be another way. It is true that these appearance states are at least in part world-directed, however, what it is for a subject to *qualify* a first-order percept as a percept, such that the subject *understands* that the world seems a certain way just is for the subject to understand that she is in the grip of an appearance which could differ from the way the world really is, and this is just to say that the subject is undergoing a conscious experience. In that regard the appearance qualifier is distinguishable. For the other qualifiers mentioned above (*e.g.,* the temporal, spatial, quantitative, &c.) merely contribute to the first-order content, adding another world-directed dimension to such content. The appearance qualifier, on the other hand, is essentially meta-representational. It does not merely add another world-directed dimension. It re-represents the relevant components of the perceptual state. Perceptions are already "appearances." To tag an appearance with an appearance qualifier is to have a representation of a visual appearance as an appearance.

It is also important to notice that while these states are higher-order, they do not target numerically distinct, extrinsic states. The conjoined first-order/higher-order components of the "qualified" belief form a complex state, which is self-directed. The quotational structure outlined in section 1 provides a precise structure for qualified perceptual belief states. And if such states are best characterized as quotational then we

have one plausible answer to the question "Why would there be mental quotation in the first place?" In particular, the answer is that mental quotation evolved early on as part of our capacity to understand the is/seems distinction, which requires attributing appearance states to oneself. Ultimately, the ability to attribute such states to others was advantageous.

While the kind of state that underwrites our understanding of the is/seems distinction might be characterizable in terms of mental quotation, one might still wonder why these qualified states would be quotational in particular. After all, that is not how Lurz describes them. But it seems reasonable to hypothesize that in rigging up the relevant components of the mind-reading system in the modular piecemeal way according to which evolution seems to proceed, nature built upon already existing structures. Thereafter to represent one of one's own states the system represents it directly, by making use of that state itself, not by means of a numerically distinct higher-order state. That is, once "seeing as" discriminations are in place, the creature who understands that seeing as discriminations entail that appearances can differ from reality has an advantage over creatures who do not understand. As stated above, one way to characterize what it is for a creature to understand the is/seems distinction is for it to have a concept of appearance, which it can then deploy to qualify beliefs. But these qualified beliefs have a certain self-directed structure, and one very efficient way to direct one's thought at one of one's own thoughts is to deploy, or embed the very target state itself to which the higher-order component is directed. This is exactly what mental quotations are alleged to do. In that regard Lurz's seems-qualifier might not be very different than Papineau's original experience operator, with the exception that Papineau thought the

experience operator operates over conscious states only. That is, it might well be the case that the same cognitive structures underwrite the kinds of phenomena that both Papineau and Lurz are trying to explain. This claim will need slight modification when one considers that the qualified beliefs described by Lurz seem to be caused by perceptual states, or states occurring "within" the visual system. If there is a clear distinction between these two, then the qualified beliefs would count as representations of numerically distinct states. However, this might not be the most accurate or theoretically fruitful way to characterize these sorts of perceptual beliefs. For it might be the case that there are mixed cognitive/perceptual states at work within the visual system. If so, then the initial qualified beliefs, which are thought to be fixed rapidly and strongly might well not target numerically distinct states. This would lend plausibility to the quotational view.

2. Future Directions

2.1 Mental Quotation, Attitudes, and Introspection

My focus in the previous chapters has been on mentally quoting perceptual states. But it is natural to wonder whether mental quotation operates over attitudes as well as perceptions. The hypothetical I want to think about is this: *if* our cognitive architecture is such that attitudes can be quoted, would such a process adequately underwrite "introspection?"

Above we captured the structure of a quotational concept by 'SEEMS<blank>'. In paradigmatic cases the blank portion of the concept is a placeholder for a perceptual

filling (‘<blue sky>’). When a perceptual state is mentally quoted, the conscious experience itself is supposed to become the content of the quoting state, and is thereby introspected. However, the crucial components of the perceptual state are thought to be exhausted by the state’s content. Attitudes, on the other hand, are by their very nature beyond contents. They’re taken *toward* contents. Thus, for the attitude to be quoted the content and attitude (type) must be quoted; to quote ‘belief that *p*’ it would not suffice to quote *p*. But attitude types are typically characterized in terms of their causal roles, *e.g.,* a belief is that kind of state which combines with desires and motivates behavior. The question that immediately arises is whether a causal role could somehow be quoted such that it becomes part of the content of the quotational state.

Consider beliefs. One way to capture the process of mentally quoting an occurrent belief is with the notion of a state having a ‘dual-role’. For example, when an occurrent belief is quoted, the lower-order belief functions as a belief (it currently figures in at least some of the relevant computational relations) *and* it functions as the content of the quoting state. In such a scenario there is no explicit attitude typing faculty that need exist. Rather, the quoting state points to and presents the attitude state itself as it is; the belief state’s first-order characteristics including its causal role or attitude type ‘speak for themselves’, as it were.

Another possibility is that the attitude types have fundamental characteristics that are somehow explicitly encoded in their instances. For example, the belief that the empire state building is taller than the Washington monument is tagged with a ‘b’ subscript. When quoted the state quoted is recognized as a belief: ‘SEEMS <(the empire state building is taller than the Washington monument)$_b$>’.

Again, the mechanism for generating quotational thoughts (the mindreading faculty) need not possess any special capacity to type attitudes, it need only latch onto some state or other and present that state, but the object state speaks for itself in that it figures in at least some of the relevant computational relations. The quotational state delegates to the object state itself, which determines its type. The type of state that is presented is determined by the object state itself. That it is a state one is currently undergoing is determined by the higher-order quoting component.

Whether or not the introspected sate is phenomenally conscious will depend in part on the first order content of the quoted stated. For example, if you think that phenomenal character is parasitic on qualitative character (as I have argued above), then assuming that attitudes do not have essential qualitative elements, then even if such states are quoted (and thereby introspected), they would not necessarily be phenomenally conscious. Thus, while at one level the subject may indeed self-attribute introspectively, at the personal conscious level the subject may not. The result fits nicely with various empirical data that support the claim that self-knowledge is *not* authoritative, and also with the anecdotal evidence that subjects "just seem to know" what they believe or don't believe without being able to articulate noninferentially why they think they believe what they do. On the other hand, if you think there is such a thing as cognitive phenomenology, then quoted attitudes might turn out to be phenomenally conscious, such that there are in fact conscious judgments/decisions. I leave this to further investigation.

2.2 Animals and Infants

Higher-order theories are sometimes charged with excluding animals and infants from being conscious.[10] There has been a lot of debate about this and I will not review that debate. To conclude this section, I just want to illustrate that the problem may not arise for the quotational in quite the same way. Above I described the process of mental quotation as a top-down procedure, but it does not have to be. One can envision the process running bottom up, such that attention would result in the seems operator getting appended to sensory contents priori to conceptualization by the quotation-generating faculty (*e.g.,* the mindreading faculty). This would have the benefit of partly enabling such creatures to navigate the is/seems distinction, but would also entail that they are phenomenally conscious. As a reminder, while mental quotations are implicitly assertoric (they implicitly assert that a seeming is underway), they do not require that the subject think a thought such as, I am now undergoing an appearance of *x.* Mental quotation is not simply linguistic quotation in inner speech.

To further develop a response to the animals and infants objection, much more would have to be said. But we can at least appreciate how the quotational model might develop a more detailed account of phenomenal consciousness for certain animals and infants.

[10] See for example, Carruthers, (2000) and Gennaro (2004)

Bibliography

Alter, T. and Walter, S. (eds.) (2006). *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* Oxford: Oxford University Press.

Armstrong, D. (1968). *A Materialist Theory of Mind.* Routledge

Armstrong, D. (1977). "The Casual Theory of Mind," *Neue Hefte Für Philosophie* 11:82-95.

Armstrong, D. (1984). "Consciousness and Causality," in D. Armstrong and N. Malcolm, *Consciousness and Causality.* Blackwell.

Ashwin, P.T. and Tsaloumas, M.D. (2007). "Complex Visual Hallucinations (Charles Bonnet Syndrome) in the Hemianopic Visual Field Following Occipital Infarction. *Journal of Neurological Sciences,* 263 (1-2):184-186.

Azzopardi, P. and Cowey, A. (1993). "Preferential Representation of the Fovea in the Primary Visual Cortex." *Nature* 361, 719-721.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B (1997). *In the Theater of Consciousness.* Oxford: Oxford University Press.

Baars, B., Ramsøy, T.Z., and Laureys, S. (2003). "Brain, Conscious Experience, and the Observing Self," *Trends In Neuroscience,* 26(12), 671-675.

Bach, K., 1987, *Thought and Reference*, Oxford: Oxford University Press.

Berti, A. and Rizzolatti, G. (1992). "Visual Processing without Awareness: Evidence from Unilateral Neglect." *Journal of Cognitive Neuroscience* 4:345-351.

Block, N.J. (1983). "Mental Pictures and Cognitive Science," *Philosophical Review, 93.*

Block, N.J. (1986). "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy*, 10: 615–678. Reprinted in Stephen P. Stich and Ted A. Warfield, eds., *Mental Representation: A Reader*, Oxford: Blackwell, 1994.

Block, N.J. (1990). "Inverted Earth," *Philosophical Perspectives,* 4: 52-79.

Block, N.J. (1995). "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18 (2): 227-88.

Block, N.J. (1996). "Mental Paint and Mental Latex," in Villanueva (1996).

Block, N.J., Flanagan, O., and Guzeldere, G. (eds.) (1997). *The Nature of Consciousness: Philosophical Debates,* Cambridge, MA: MIT/Bradford.

Block, N.J (2003). "Mental Paint." In Martin Hahn and Bjorn Ramberg, eds., *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, 165-200. Cambridge, Mass.: The MIT Press, 2003. Reprinted in Block 2007, 533-563; page references are to the reprinted version.

Block, N.J. (2007). *Consciousness, Function, and Representation*. Cambridge, Mass.: MIT Press.

Block, N.J. (2011) "The Higher-Order Approach to Consciousness is Defunct," *Analysis,* 71(3):419-31.

Botterill, G. and Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge: Cambridge University Press.

Burge, T. (1979). "Individualism and the Mental," in French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, pp. 73–121.

Burge, T. (1986). "Individualism and Psychology," *Philosophical Review*, 95: 3–45.

Byrne, Alex (1997), 'Some Like it HOT: Consciousness and Higher-Order Thoughts', *Philosophical Studies,* 86, pp. 103-29.

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

Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory.* Cambridge: Cambridge University Press.

Carruthers, P. (2005). *Consciousness: Essays from a Higher-Order Perspective.* Oxford: Clarendon.

*Carruthers, P. (2009). "How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition," *Behavioral and Brain Sciences* 32 (2009).

Carruthers, Peter, "Higher-Order Theories of Consciousness", *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/consciousness-higher/>

Chalmers, D. (1996). *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, D. (2002). *Philosophy of Mind: Classical and Contemporary Readings.* New York: Oxford University Press.

Chalmers, David (2006), "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, eds. T. Alter and S. Walter (Oxford University Press).

Crane, T. 1991, "All the Difference in the World," *Philosophical Quarterly*, 41: 1–25.

Clark, H.H., and Clark, E.V. (1977). *Psychology and Language*. New York: Harcourt Brace Jovanovich.

Crick, F. and Koch, C. (1990). "Towards a Neurobiological Theory of Consciousness." *Seminars in the Neurosciences,* 2.

Davidson, D. (1979). "Quotation," *Theory and Decision* 11:27-40.

Davies, M. and Humphreys, G. (1993) eds., *Introduction to Consciousness: Psychological and Philosophical Essays,* Blackwell.

DeBellis, M. (1991). "The Representational Content of Musical Experience," *Philosophy and Phenomenological Research* 51:303-24.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., DehaeneLambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). "Imaging unconscious semantic priming." *Nature*, 395, 597-600.

Dehaene, S. and Changeux, J.P., (2011). "Experimental and Theoretical Approaches to Conscious Processing." *Neuron,* 70 (2):200-227.

Dennett, D. (1978). Why You Can't Make a Computer that Feels Pain, in *Brainstorms*. Cambridge, MA: MIT Press: 190-229.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. (1991) *Consciousness Explained*, (Little, Brown and Company).

Dennett, D. (1992) 'The Self as the Center of Narrative Gravity', *Self and Consciousness:Multiple Perspectives*, Kessel, Cole, Johnson (eds.), (Erlbaum).

Dennett, D. and Kinsbourne, M. (1995a). 'Time and the observer: The where and when of consciousness in the brain', *Behavioral and Brain Sciences* XV, (2), pp. 183-247.

Dennett, D. and Kinsbourne, M., (1995b). 'Escape from the Cartesian Theater', *Behavioral and Brain Sciences*, XV, pp. 234-46.

Dennett, D. (2001), 'Are We Explaining Consciousness Yet?', *Cognition*, LXXIX, pp. 221-37.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., and Saxe, R., (2011). "fMRI Item Analysis in a Theory of Mind Task." *Neuroimage* 55:705-712.

Doricchi, F.., and Galati, G. (2000). "Implicit Semantic Evaluation of Object Symmetry and Contralesional Visual Denial in a Case of Left Unilateral Neglect with Damage of the Dorsal Paraventricular White Matter. *Cortex.* 36:337-350.

Dretske, F. (1981). *Knowledge and the Flow of Information.* Cambridge: MIT Press/Bradford Books.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes,* Cambridge: MIT Press/Bradford Books.

Dretske, F. (1995). *Naturalizing the Mind.* Cambridge: MIT Press.

Driver, J., and Mattingly, J.B. (1998). "Parietal neglect and visual awareness. *Nature Neuroscience,* 1:17-22.

Driver, J., and Vuilleumier, P. (2001). "Perceptual Awareness and its Loss in Unilateral Neglect and Extinction." *Cognition,* 79:39-88.

Feinberg, T. (2000), "The Nested Hierarchy Theory of Consciousness: A neurobiological solution to the problem of mental unity, *Neurocase* 6:75-81.

Feinberg, T. (2001). *Altered Egos: How The Brain Creates the Self.* New York: Oxford University Press.

Feinberg, T. (2009). *From Axons to Identity: Neurological Explorations of the Nature of the Self.* New York: W.W. Norton.

Fodor, J. (1975). *The Language of Thought,* Cambridge: Harvard University Press.

Fodor, J. (1987). *Psychosemantics.* Cambridge, MA.: MIT Press.

Fodor, J. (1990). "Psychosemantics or Where do Truth Conditions Come From?" in Lycan (1990), pp. 312-38.

Fodor, J. (1991). "Replies," in Loewer and Rey (1991), pp. 255-319,

Fodor, J. (1995) *The Elm and the Expert: Mentalese and its Semantics*, Cambridge, MA: MIT Press.

Frege, G. (1892/1976) *Philosophical and Mathematical Correspondence.* In M. Beaney (Ed.) *The Frege Reader,* Blackwell.

Garcia, J.A., and R.A., Koelling (1967), 'A Comparison of Aversions Induced by X rays, Toxins, and Drugs in the Rat', *Radiation Research Supplement*, VII, pp. 439-50.

Geach, P. (1957). *Mental Acts,* London: Routledge & Kegan Paul.

Gennaro, R. (1996). *Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness.* Amsterdam: John Benjamins.

Gennaro, R. (2004). "Higher-Order Thoughts, Animal Consciousness, and Misrepresentation: A reply to Carruthers and Levine. In R.J. Gennaro (Ed.) *Higher-Order Theories of Consciousness.* Amsterdam and Philadelphia: John Benjamins.

Gennaro, R. (2006), "Between Pure Self-Referentialism and the Extrinsic HOT Theory of Consciousness," in *Self-Representational Approaches to Consciousness*, eds. U. Kriegel and K. Williford (MIT Press).

Gertler, B. (2006). "Consciousness and Qualia Cannot be Reduced," in R. Stainton (ed) *Contemporary Debates in Cognitive Science* Blackwell.

Godfrey-Smith, P. (1994). "A Continuum of Semantic Optimism," in S. Stich and T.

Warfield (eds), *Mental Representation,* Oxford: Blackwell, pp. 259-77.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature,* Cambridge: Cambridge University Press.

Goldman, A. (1993). "Consciousness, Folk Psychology, and Cognitive Science." *Consciousness and Cognition* 2: 364-383.

Hardcastle, V. (1999). *The Myth of Pain.* Cambridge, MA: MIT Press.

Hardcastle, V. (2008). "The Binding Problem," in W. Bechtel and G. Graham (eds.) (1998/2008), *A Companion to Cognitive Science,* Blackwell.

Hardin, C.L. (1988). *Color for Philosophers: Unweaving the Rainbow.* Indianapolis: Hackett.

Harman, G. (1982). "Conceptual Role Semantics," *Notre Dame Journal of Formal Logic,* 23:242-56.

Harman, G. (1990). "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4:31-52.

Harms, W. F., (1998), "The Use of Information Theory in Epistemology", *Philosophy of Science*, 65(3): 472–501.

Hellie, B. (2007). "Higher-Order Intentionality and Higher-Order Acquaintance," *Philosophical Studies,* 134: 289-324.

Horgan, T. and Tienson, J. (2002). "The Intentionality of Phenomenology and The Phenomenology of Intentionality." In Chalmers, D. (ed) (2002). *Philosophy of Mind: Classical and Contemporary Readings.* New York: Oxford University Press.

Jehle, D., and Kriegel, U. (2004). "An Argument Against Dispositionalist HOT Theory," *Philosophical Psychology*, Vol. 19, No. 4, pp. 463-476.

Jordon, G. and Mollon, J.D. (1993). "A Study of Women Heterozygous for Color Deficiencies," *Vision Research* 33: 1495-1508.

Jordon, G., Deeb, S.S., Bosten, J.M., and Mollon, J.D. (2010). "The Dimensionality of Color Vision in Carriers of Anomalous Trichromacy." *Journal of Vision,* 10(8):

Kim, J. (1992). "Multiple Realization and the Metaphysics of Reduction," *Philosophy and Phenomenological Research* 52:1-26.

Kim, J. (1998). *Mind in a Physical World,* MIT Press.

Kirk, R. (1994). *Raw Feeling.* Oxford: Oxford University Press.

Kobes, B.W. (1995). "Telic Higher-Order Thoughts and Moore's Paradox." *Philosophical Perspectives* 9: 291-312.

Kriegel, U. (2002). "PANIC theory and the Prospects for a Representational Theory of Phenomenal Consciousness," *Philosophical Psychology,* 15 (1):55-64.

Kriegel, U. (2005). "Naturalizing Subjective Character." *Philosophy and Phenomenological Research*, Vol, 71, No. 1, pp. 23-57.

Kriegel, U. (2006). "The Same-Order Monitoring Theory of Consciousness." In U. Kriegel, and K. Williford (eds.), *Self-Representational Approaches to Consciousness*. Cambridge: MIT Press.

Kriegel, U. (2009). *Subjective Consciousness,* Oxford: Oxford University Press.

Kriegel, U. and Williford, K. (eds.) (2006). *Self Representational Approaches to Consciousness.*Cambridge, MA: MIT Press.

Kripke, S. (1972). *Naming and Necessity*, Oxford: Blackwell.

Kripke, S. (1979). "A puzzle about belief," in *Meaning and Use*, edited by A. Margalit. Dordrecht and Boston: Reidel.

Lamme, V.A. (2004). "Separate Neural Definitions of Visual Consciousness and Visual Attention: A Case for Phenomenal Awareness." *Neural Networks* 17 861-872.

Lamme, V.A. (2006). "Towards a True Neural Stance on Consciousness." *Trends in Cognitive Science.* 10, 494-501.

Lau, H. (2008). "A Higher-Order Bayesian Decision Theory of Perceptual Cosnciousness" in R. Banjeree and B.K. Chakrabarti (eds.) *Progress in Brain Research,* Vol. 168.

Lau, H. and Brown, R. (forthcoming). "The Emperor's New Phenomenology: The Empirical Case of Conscious Experience without First-Order Representations," in A. Pautz and D. Stoljar (eds) *Festschrift for Ned Block*, Cambridge: MIT Press.

Lau, H. and Rosenthal, D. (2011). "Empirical Support for Higher-Order Theories of Conscious Awareness," *Trends in Cognitive Sciences,* 15(8):365-373.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical* Quarterly 64:354-61.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness.* Oxford and New York: Oxford University Press.

Levine, J. (2006). "Conscious Awareness and (Self-) Representation," in U. Kriegel, and K.

Williford (eds.), *Self-Representational Approaches to Consciousness.* Cambridge: MIT Press.

Lewis, D. (1983) *Philosophical Papers,* Vol. 1. New York: Oxford University Press.

Lewis, D. (1988). "What Experience Teaches." *Proceedings of the Russellian Society.*

Reprinted in Chalmers (2002).

Loar, B. (1981). *Mind and Meaning.* New York: Cambridge University Press.

Loar, B. 1988, "Social Content and Psychological Content," in R. Grimm and D. Merrill (eds.), *Contents of Thought*, Tucson: University of Arizona Press.

Loar, B. (2003). "Transparent Experience and the Availability of Qualia." In Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Perspectives.* Oxford: Oxford University Press.

Lopes, D.M.M. (2000). "What is it Like to See with your Ears?" The Representational Theory of Mind," *Philosophy and Phenomenological Research,* 60:439-53.

Lycan, W. (1987). *Consciousness.* Cambridge, MA: Bradford Books/MIT Press.

Lycan, W. (1990). *Mind and Cognition.* Oxford: Blackwell.

Lycan, W. (1990/1997). "Consciousness as Internal Monitoring," in Block, N.J., Flanagan, O., and Guzeldere, G. (eds.) (1997). *The Nature of Consciousness: Philosophical Debates,* Cambridge, MA: MIT/Bradford.

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

Lycan, W. (2001). "A Simple Argument for a Higher-Order Representation Theory of Consciousness," *Analysis*, 61: 3-4.

Lycan, W. (2008). "The Superiority of HOP to HOT" In William Lycan & Jesse J. Prinz (eds.), *Mind and Cognition*. Blackwell.

Lycan, W. and Pappas, G. (1972). What Is Eliminative Materialism? *Australasian Journal of Philosophy* 50:149-59.

Macknik, S.L. and Martinez-Conde, S. (2007). "The Role of Feedback in Visual Masking

and Visual Processing." *Advances in Cognitive Psychology,* 3(1-2):125-152.

Marcel, A. (1998). "Blindsight and Shape Perception: deficit of visual consciousness or of visual function?" *Brain,* 121.

Marshall, J.C., and Halligan, P.W. (1988). "Blindsight and Insight in Visiospatial Neglect." *Nature.* 336:766-767.

Maund, B. (1995). *Colors: Their Nature and Representation.* Cambridge: Cambridge University Press.

von der Malsburg, C. (1981) "The Correlation Theory of Brain Function," Technical Report 81-2, Max-Planck-Institute for Biophysical Chemistry, Gottingen.

McGinn, C. (1977). "Charity, interpretation, and belief," *Journal of Philosophy*, 74: 521–535.

McGinn, C. (1982).*The Character of Mind.* Oxford U. Press.

McGinn, C. (1989). "Can We Solve the Mind Body Problem?" in N. Block, O. Flanagan, and G.

Guzeldere (eds.), *The Nature of Consciousness,* Cambridge, MA: Bradford Books/MIT Press, 529-542.

Meyer, D.E., and Schvaneveldt, R.W. (1971). "Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations." *Journal of Experimental Psychology, 90, 227-234.*

Milikan, R. (1984). *Language, Thought, and Other Biological Categories,* Cambridge: MIT Press/Bradford Books.

Milner, P. (1974). "A Model for Visual Shape Recognition," *Psychological Review.* 81:521-535.

Milner, D. and Goodale, M. (1995). *The Visual Brain in Action.* Oxford: Oxford University Press.

Moore, G.E. (1903). "The Refutation of Idealism," in Moore's *Philosophical Papers,* London: Routledge and Kegan Paul.

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton: Princeton University Press.

Nagel, T. (1974). "What is it Like to Be a Bat?" *Philosophical Review*. 83: 435-450.

Natsoulas, T. (1993). "What is Wrong with Appendage Theory of Consciousness." *Philosophical Psychology* 6: 137-154.

Natsoulas, T. (1999). "The Case for Intrinsic Theory: III. Intrinsic Inner Awareness and the Problem of Straightforward Objectification." *Journal of Mind and Behavior* 19: 1- 20.

Neander, K. (1991). "Functions as Selected Effects" *Philosophy of Science,* 58:168-84.

Neely, J. H. (1976). "Semantic priming and retrieval from lexical memory: Evidence for facilitory and inhibiting processes." *Memory and Cognition*, 4, 648-654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106, 226-254. Nelkin, N. (1989). "Propositional Attitudes and Consciousness." *Philosophy and Phenomenological Research,* 49: 413-30.

Newton, J.R., and Eskew, R.T., Jr. (2003). "Chromatic Detection and Discrimination in the Periphery: a postreceptoral loss of color sensitivity." *Vision Neuroscience* 20, 511-521.

O'Regan, J. K., and A. Noë (2001) "A Sensorimotor Approach to Vision and Visual Consciousness" *Behavioral and Brain Sciences*, 24, pp.883–975.

Papineau, D. (1987). *Reality and Representation.* Oxford: Blackwell.

Pautz, A. and Stoljar, D. (eds) (forthcoming). *Festschrift for Ned Block*, Cambridge: MIT Press.

Peacocke, C. (1983). *Sense and Content.* Oxford University Press.

Picciuto, V. and Carruthers, P. (in press). "Inner Sense," in S. Biggs, M. Matthen, and D. Stokes (eds) *Perception and its Modalities,* Oxford University Press.

Pitt, D. (2004). The phenomenology of cognition Or What is it like to think that P? *Philosophy and Phenomenological Research*, 69: 1-36.

Prinz, J. (2007). "Mental Pointing: phenomenal knowledge without concepts," *Journal of Consciousness Studies*. 14 (9-10):184-211.

Putnam, Hilary (1975). "The Meaning of Meaning," *Philosophical Papers, Vol. II : Mind, Language, and Reality*, Cambridge: Cambridge University Press.

Quine, V.V.O. (1960). *Word and Object.* Cambridge, Mass: MIT Press.

Rahnev, D., Maniscalo, B., Graves, T., Huang, E., de Lange, F., Lau, H., (in press). "Attention Induces Conservative Subjective Biases in Visual Perception." *Nature Neuroscience.* 14, 1513-1515.

Ramachandran, V. and Blakeslee, S. (1998). *Phantoms in the Brain.* Fourth Estate.

Recanti, F. (2004). *Literal Meaning.* Cambridge University Press.

Rees, G., Kreiman, G., and Koch, C. (2002). "Neural Correlates of Consciousness in Humans." *Nature Reviews. Neuroscience.* 3(4):261-270.

Reichenbach, H. (1947). *Elements of Symbolic Logic,* New York: Macmillan.

Rey, G. (1982). “A Reason for Doubting the Existence of Consciousness,” In Richard J. Davidson, Sophie Schwartz & D. H. Shapiro (eds.), *Consciousness and Self-Regulation, Vol. 3*. New York: Plenum

Rey, G. (1992). “Sensational Sentences Switched”, *Philosophical Studies* 67: 73–103.

Rey, G (1993). “Sensational Sentences” in *Consciousness,* M. Davies and G. Humphrey (eds.), Oxford, UK: Basil Blackwell, pp. 240–57.

Rey, G. (1988). “A Question About Consciousness,” in Block *et al* (1997).

Rey, G. (1998). “A Narrow Representationalist Account of Qualitative Experience,” *Noûs*, Vol. 32, Supplement: Philosophical Perspectives, 12, Language, Mind, and Ontology, pp. 435-457

Rey, G. (2008). “(Even Higher-Order) Intentionality Without Consciousness. *Revue Internationale de Philosophie*, 243.

Rosenthal, D. (1990). “A Theory of Consciousness.” Report No. 40/1990 on MIND and BRAIN. *Perspectives in Theoretical Psychology and the Philosophy of Mind* (ZiF). University of Bielefeld.

Rosenthal, D. (1993). “Higher-Order Thoughts and the Appendage Theory of Consciousness.” *Philosophical Psychology* 6: 155-166.

Rosenthal, D. (2000). “Consciousness and Metacognition,” In D. Sperber (ed.), *Metarepresentation: Proceedings of the Tenth Vancouver Cognitive Science Conference.* New York: Oxford University Press.

Rosenthal, D. (2002). “Explaining Consciousness.” In *Philosophy of Mind: Classical and Contemporary Readings,* (ed) D. Chalmers, 406-21. Oxford: Oxford University Press.

Rosenthal, D. (2002b). "How Many Kinds of Consciousness." *Consciousness and Cognition,* 11(4) 653-665.

Rosenthal, D. (2005). *Consciousness and Mind.* Oxford: Clarendon.

Rosenthal, D. (2009). "Higher-Order Theories of Consciousness." In *Oxford Handbook of the philosophy of Mind,* (eds.) B. McLaughlin and A. Beckermann, 239-252. Oxford: Clarendon Press.

Russell, B. (1940/2013) *An Inquiry Into Meaning and Truth,* Routledge.

Segal, G. (2000). *A Slim Book about Narrow Content*, Cambridge: MIT Press.

Sellars, W. (1963). *Science, Perception, and Reality,* London: Routledge and Kegan Paul.

Sellars, W. (1967). *Science and Metaphysics,* London: Routledge and Kegan Paul.

Sellars, W. (1971). "Science, Sense Impressions and Sensa: A Reply to Corman" *Review of Metaphysics,* 24, 391-447.

Sellars, W. (1975), "The Adverbial Theory of the Objects of Sensation," *Metaphilosophy,* 6:144-160.

Shoemaker, S. (1994a). Phenomenal Character, *Nous,* 28: 21-28

Shoemaker, S. (1994b). Self-knowledge and inner sense. Lecture III: The phenomenal character of experience, *Philosophy and Phenomenological Research*, 54: 291-314.

Shoemaker, S. (2002). Introspection and Phenomenal Character. In D.J. Chalmers (ed.), *Philosophy of Mind.* Oxford and New York: Oxford University Press.

Siewart, C. (1998). *The Significance of Consciousness.* Princeton: Princeton University Press.

Silvanto, J. and Rees, G. (2011). "What Does Nueral Plasticity Tell us About the Role of

Primary Visual Cortex (V1) in Visual Awareness?" *Frontiers in Psychology* 2:6. doi:10.3389/fpsyg.2011.00006.

Smart, J.J.C. (1959). "Sensations and Brain Processes," *Philosophical Review* 68:141-56.

Sperber, D. and Wilson, D. (2002). "Pragmatics, Modularity and Mind Reading," *Mind and Language,* 17 (1-2): 3-23.

Stampe, D. (1977). "Towards a Causal Theory of Linguistic Representation," *Midwest Studies in Philosophy,* Minneapolis: University of Minnesota Press, pp. 42-63.

Strawson, G. (1994). *Mental Reality.* Cambridge, MA: MIT Press.

Tarski, A. (1933) "The Concept of Truth in Formalized Languages," reprinted in Tarski, *logic, Semantics, Metamathematics,* 2nd edn. (Indianapolis: Hackett, 1983), 152-278.

Treisman, A., and Schmidt, H. (1982). "Illusory Conjunctions in the Perception of Objects," *Cognitive Psychology,* 14:107-141.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247, 301–306.

Tulving, E., Schacter, D., & Stark, H. A. (1982). "Priming effects in word-fragment completion are independent of recognition memory." *Learning, Memory and Cognition*, 8, 336–341.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Tye, M. (1995a). "A Representational Theory of Pains and their Phenomenal Character," in Block, Flanagan, and Guzeldere (1997).

Tye, M. (2000). *Consciousness, Color, and Content.* Cambridge, MA: MIT Press.

Tye, M. (2006) "Non-conceptual Content, Richness, and Fineness of Grain," in

Perceptual Experience, (eds.) Szabo Gendler, T. and Hawthorne, J., Oxford: Oxford University Press.

Van Gulick, R. (2004). "Higher-Order Global States (HOGS): An Alternative Higher-Order Model Of Consciousness." In R. Gennaro (ed.) *Higher-Order Theories of Consciousness*. Philadelphia: John Benjamins.

Van Gulick, R. (2006). "Mirror Mirror – Is That All?" In U. Kriegel, and K. Williford (eds.), *Self-Representational Approaches to Consciousness*. Cambridge: MIT Press.

Villanueva, E. (ed.) (1996). *Philosophical Issues, 7: Perception*. (Atascadero, CA: Ridgeview Publishing.)

Vinueza, A. (2000). "Sensations and the Language of Thought." *Philosophical Psychology* 13(3): 373–392.

Von der Malsburg, C. (1981). "The Correlation Theory of Brain Function." Technical Report 81-2, Max-Planck-Institute for Biophysical Chemistry, Gottingen.

Weisberg, J. (2011). "Abusing the Notion of What-It's-Likeness: A Response to Block." *Analysis* 71 (3):438-443.

Weisberg, J. (2010). "Misrepresenting Consciousness." *Philosophical Studies.* Advance Access published 12 May 2010, doi:10.1007/s11098-010-9567-3.

Weisberg, J. (2008), 'Same Old, Same Old: The Same-Order Representation theory of Consciousness and the Division of Phenomenal Labor' *Synthese* 160 (2):161-81.

Weiskrantz, L. (1986). *Blindsight.* Oxford, England: Oxford University Press.

Weiskrantz, L. (1997). *Consciousness Lost and Found.* Oxford: Oxford University Press.