Philos Stud (2011) 152:361–383
DOI 10.1007/s11098-009-9444-0

Phenomenal consciousness with infallible self-representation

Chad Kidd

Published online: 26 November 2009

Abstract In this paper, I argue against the claim recently defended by Josh Weisberg that a certain version of the self-representational approach to phenomenal consciousness cannot avoid a set of problems that have plagued higher-order approaches. These problems arise specifically for theories that allow for higher-order misrepresentation or—in the domain of self-representational theories—self-misrepresentation. In response to Weisberg, I articulate a self-representational theory of phenomenal consciousness according to which it is *contingently* impossible for self-representations tokened in the context of a conscious mental state to misrepresent their objects. This contingent infallibility allows the theory to both acknowledge the (logical) possibility of self-misrepresentation and avoid the problems of self-misrepresentation. Expanding further on Weisberg's work, I consider and reveal the shortcomings of three other self-representational models—put forward by Kreigel, Van Gulick, and Gennaro—in order to show that each indicates the need for this sort of infallibility. I then argue that contingent infallibility is in principle acceptable on *naturalistic* grounds *only if* we attribute (1) a neo-Fregean kind of directly referring, indexical content to self-representational mental states and (2) a certain ontological structure to the complex conscious mental states of which these indexical self-representations are a part. In these sections I draw on ideas from the work of Perry and Kaplan to articulate the context-dependent semantic structure of inner-representational states.

Keywords Phenomenal consciousness · Higher-order representationalism · Self-representationalism · Infallibility · Naturalism · Indexical content

C. Kidd (✉)
University of California, Irvine, 222 HOB2, 625 Mesa Road, Irvine, CA 92697-4555, USA
e-mail: ckidd@uci.edu

1 Introduction: self-representational and higher-order representational theories of consciousness

The basic thesis of a self-representational approach to consciousness is that a conscious mental state represents both the world and itself. According to such views, all conscious mental states have a two-part representational structure that consists of a world-directed representation and a self-directed representation—what I will call, respectively, an *outer*- and *self*-representation. The outer-representation represents some feature of the world external to the mental state of which it is a part; the so-called self-representation, on the other hand, represents either itself or some other part of the comprehensive mental state it helps compose. This kind of view has been motivated by a desire to maintain key features of the higher-order representational approach to consciousness while avoiding what some think are insurmountable difficulties that jeopardize all higher-order models. The basic difference between the higher-order and self-representational approaches to consciousness is that the higher-order representational approaches hold that phenomenally conscious experience is the product of two mental states—not just one, as on the self-representational approach—one of which represents the phenomenal content of the other. Higher-order theories thus hold that there is no need to attribute a complex self-representational structure to conscious states.

The virtues of the higher-order approach are many. I here mention only two, each of which many self-representational theorists have sought to preserve in their own theories. First, the higher-order approach provides the means to attain a non-eliminative, naturalistic explanation of phenomenal consciousness.[1] According to higher-order models, phenomenal consciousness is the product of the representation of one mental state by another. So, by hypothesis, it seems that the naturalistic theory of mental representation is the key to unlock the mysteries of both intentionality and phenomenal consciousness, grounding both of these phenomena in the natural world. Of course, this colossal step forward in the philosophy of mind is still just a philosopher's reverie; for despite all the efforts of some of our best philosophers of mind, we still have not uncovered an adequate naturalistic metaphysical basis for mental representation.[2] But for those who are still hopeful, this is a very attractive feature of higher-order views.

Second, higher-order theories provide a framework that grounds the intuition that phenomenal consciousness consists in, or at least essentially includes, a consciousness of one's own mental life.[3] This intuition has been shared by many philosophers throughout the centuries, from Aristotle, to Descartes, to Brentano and his student

[1] For a discussion of this explanatory method see (Rosenthal 2002).

[2] For a review of the variety of naturalistic projects and sympathetic discussion of the inadequacies in each see (Loewer 1997). For a presentation of the problems facing any naturalizing project that attempts to reduce representation to a causal relation, even with added bells and whistles, see Smith (1999) and Bilgrami and Rovane (2005).

[3] Virtually all higher-order and self-representational theorists acknowledge and attempt to explain the intuition of self-awareness. The idea that phenomenal consciousness includes self-awareness has come under attack in work by Dretske (1995, 2002), Tye (1995, 2000), and Thomasson (2000, 2005, 2006).

Husserl[4]; and, in the contemporary literature, it has been articulated and defended in many different ways.[5]

This second virtue, however, also contains the germ of what many self-representational theorists see as insurmountable difficulties for higher-order theories: problems that arise for theories that assimilate the structure of mind-directed mental representations to that of world-directed mental representations. World-directed or 'outer'-representations can misrepresent the objects they are about. And so, *mutatis mutandis*, mind-directed or 'self'-representations can misrepresent the objects they are about. Now, in some cases, this assimilation is a welcome result. We do, in fact, often seem to have a tenuous and epistemically unreliable hold on our own mental lives,[6] and we want a theory of phenomenal consciousness to reflect and explain this fact. But there are also cases of mind-directed representation where it seems that misrepresentation should be excluded altogether or, at least, limited. So, in short, the problems for higher-order theories stem from the need to reconcile the assimilation of the structure of self- and outer-representation with the idea that self-misrepresentation should be excluded or limited.

This and other difficulties of higher-order misrepresentation have been well developed in the literature,[7] and I will present the relevant problems in sufficient detail in the next section. It has been noticed recently that these same sorts of problems also raise a proportionally devastating threat to a certain vintage of self-representational theories of consciousness, in particular, *naturalistic* self-representational theories.[8] And so, it seems, as Weisberg has concluded in a recent paper on this issue, "SORT (self-representationalism) is not in a better explanatory position than HORT (higher-order representationalism)" (Weisberg 2008, p. 179).

My contribution to this chain of insights is to show, contrary to Weisberg, that self-representationalism has the resources to maintain the explanatory edge over higher-order representationalism because there is still *one* naturalistic self-representational theory that has the resources to avoid the problems that arise from self-misrepresentation—in particular, a self-representational theory according to which self-representation is *infallible*.

I will show how every other naturalistic self-representational theory of consciousness besides the one I defend falls prey to the problems of self-misrepresentation. I will then show how a certain model of the *logical structure* of the representational content of self-representation and a certain (now commonly used) self-representational model of the *ontological structure* of a conscious mental state together provide for a type of self-representational infallibility that is consistent with a naturalistic metaphysical basis for representation in general.

[4] See Caston (2002) for an interpretation of Aristotle's view of consciousness; see Smith (2004b) for a discussion of Descartes in this regard; and see passages in Brentano (1995, Book Two Chap. II, III) and Husserl (1962, Sect. 45).

[5] For a defense and further developments of the idea that self-awareness is an essential part of phenomenal consciousness from a phenomenological perspective see Smith (2004a) and Kriegel (2009).

[6] See the enlightening discussion of this point in (Schwitzgebel 2008).

[7] See Byrne (1997), Neander (1998), Seager (1999, Chap. 3), Levine (2001, Chap. 4), Van Gulick (2004, 2006), and (Kriegel 2006).

[8] See Levine (2006) and Weisberg (2008).

2 The problems of higher-order misrepresentation

Two problems arise for higher-order representational theories from the possibility of higher-order misrepresentation. On one hand, it seems that the common or traditional notion of phenomenal quality cannot be reconciled with the possibility of misrepresenting the phenomenal qualities of a world-directed or 'outer'-representational state.[9] This is so because, according to the traditional notion of phenomenal quality, having an experience with a blue phenomenal quality entails having a special type of awareness—a 'background' or 'prereflective' awareness[10]—of a blue phenomenal quality, and, conversely, having a prereflective awareness of a blue phenomenal quality entails having that experience. So, according to the traditional notion of phenomenal quality, the existence of a phenomenal quality in a subject's mind and a subject's (prereflective) awareness of that phenomenal quality mutually entail one another. But the possibility of higher-order misrepresentation breaks up this mutual entailment relation in both directions. If a higher-order mental state can misrepresent the phenomenal qualities of the first-order state it is about, then having an experience with a blue phenomenal quality no longer entails that one also has a prereflective awareness of a blue phenomenal quality. And, since the prereflective awareness provided by the higher-order representation could be radically falsidical, having a prereflective awareness of a blue phenomenal quality does not entail that one also has an experience with that quality.[11]

The other problem is that the possibility of higher-order misrepresentation puts the explanatory power of higher-order representationalism in jeopardy. As was mentioned, one of the virtues of this theory is that it provides for a reductive/naturalistic but non-eliminative explanation of phenomenal consciousness in terms of mind-directed representational states. But if it is possible to have an awareness of an experience with blue phenomenal qualities without actually having such an experience tokened in one's mind, then it seems the production of the phenomenal blueness for the subject in such cases would be due to the higher-order mental state *alone,* and not a representation relation between two mental states. As Alex Byrne says about such cases of higher-order misrepresentation with reference to David Rosenthal's higher-order thought theory of consciousness,

[9] See Neander (1998) and Levine (2001, pp. 108, 168).

[10] Brentano calls this special kind of awareness of the phenomenal qualities of an experience "inner-consciousness" (Brentano 1995, p. 128). He describes it as a consciousness "given alongside" the consciousness of the object of the world-directed experience. Husserl made this distinction by saying that there is an type of awareness of our mental lives that is part and parcel of "living through" an experience, and that this kind of awareness is phenomenologically distinct from a reflective awareness of experience (Husserl 1962, Sect. 45, 1991, Appendix VIII); see also Zahavi (1998). Sartre followed suit, and was first to call it "prereflective" self-awareness (Sartre 1956, Introduction, Sect. III). More recently, David W. Smith has isolated this phenomenon in the framework of a self-representational theory of consciousness, calling it "inner-awareness" (Smith 1986, 1989, 2005). Zahavi has also done much to bring the phenomenological peculiarities of prereflective self-awareness to light for contemporary philosophers of mind (Zahavi 1999, 2005). Also see the discussion of the consciousness of consciousness as a part of the background or margin of conscious experience in Gurwitsch (1985a).

[11] For detailed exposition of this problem for higher-order theories, see Neander (1998) and Levine (2001, p. 168).

> the strategy behind the higher-order thought hypothesis…is this. We start with an account of mental states that does not presuppose that they are conscious. We say that a mental state is to have (underived) intentionality or sensory properties (or both)…. We then say that to be a conscious mental state is to be the object of another mental state…. And if this is correct, then we have shown how to construct all conscious states from entirely nonconscious building blocks. The present problem is that if the higher-order thought hypothesis is true, higher-order thoughts that one is in a sensory state, and which occur in the right way, must be alone sufficient for phenomenal consciousness. (Byrne 1997, p. 122)

The same consequence follows for any other higher-order representation theory that acknowledges the possibility of higher-order misrepresentation, and this renders them incompatible with the reductive explanatory strategy they seek to employ.

The higher-order representational theorist could avoid the first problem—concerning the incompatibility of higher-order misrepresentation and the traditional notion of phenomenal quality—by stipulating that he is working with a *new* concept of 'phenomenal quality'; to wit, one according to which a mutual entailment relation between having experiences with phenomenal qualities and (prereflective) awareness of these phenomenal qualities does not obtain.[12] To go this way, however, is really only to give up on the task of trying to construct a representational theory of consciousness that provides an account of the traditional notion of phenomenal quality while trying to placate concerns about this strategy by calling both the new and old concepts by the same name—a strategy that to my mind has become unacceptably common in the treatment of many traditional issues in the philosophy of mind.[13] Moreover, regardless of whether the employment of new versions of traditional concepts is itself warranted in certain cases it seems that any theory of phenomenal consciousness that can carry out a reductive but non-eliminative explanation of the traditional notion of phenomenal quality would be a more parsimonious option than any theory that must resort to revisions.

3 The problems of self-misrepresentation: the inadequacies of three self-representational models[14]

In this section, I will examine three self-representational models of phenomenal consciousness that can be understood as answers to the two problems of higher-order representationalism presented above. We will see that, even though they can provide answers to some of these problems, each nevertheless falls prey to other problems that are much like those enumerated above.

[12] See Rosenthal (1991) for such an account.

[13] For more on this strategy, especially in the literature on consciousness, see the discussion of what Galen Strawson calls "looking glassing" a concept—i.e., the act of surreptitiously trading the traditional meaning of a term for one that is nonsensically out of sync with its traditional usage—in Strawson (2005).

[14] These problems, as they are for self-representational theories in particular, were first brought to my attention by Weisberg (2008). In that article he calls them the "problems of phenomenal intimacy."

I will also examine how these theories bring the logical structure they attribute to self-representation together with the ontological structure of conscious mental states. I examine this because an understanding of misrepresentation necessarily involves an understanding of both the logical structure of the (mis)representation and the ontological structure of the context and object of representation. So a theory of consciousness cannot provide answers to the problems of self-misrepresentation without at least tacitly assuming models of the logical structure of 'self'-directed representations and the ontological structure of conscious mental states. My conclusion in this regard is that the three models do not provide adequate answers to the problems of self misrepresentation because they do not delve deeply enough into these two matters.

In a recent article, Uriah Kriegel has brought the importance of the ontological structure of conscious acts to the fore. In this article, he presents a critical overview of many of the ontological models of conscious mental states presented or at least implicitly assumed in much of the recent literature. Of those he reviews, the ontological model that Kriegel finds the most plausible is

> ($SOMT_{10}$) A mental state M of a subject S is conscious iff (i) M^* is a (proper) part of M, (ii) $M^{\diamond}$ is a (proper) part of M, (iii) M^* is a representation of $M^{\diamond}$, and (iv) M is a *complex* of M^* and $M^{\diamond}$. (Kriegel 2006, p. 151)

A complex is a kind of mereological whole, which Kriegel distinguishes from mereological sums in the following way:

> A complex is a sum whose parts are essentially interconnected, or bound, in a certain way. The interconnection between these parts is an existence condition of a complex, but not of a sum. Thus, a molecule is a complex of atoms rather than a sum of atoms, since for the atoms to constitute a molecule they *must* be interconnected in a certain way. So whereas for a sum to go out of existence, it is necessary that one of its parts go out of existence, this is not the case with a complex. A complex can go out of existence even when its parts persist, provided that the relationship or connection among them is destroyed. (Kriegel 2006, p. 150)

Thus in order for two mental states to compose a complex mental state, they must share some special relation R to one another, and R "must be (i) an existence (and identity) condition of the whole, but (ii) neither an existence condition nor an identity condition on any of the parts" (Kriegel 2006, p. 152).

In other words, the ontological structure of a conscious mental state according to Kriegel's $SOMT_{10}$ is such that the self-representational mental state and the outer-representational mental state are *ontologically independent* of one another—i.e., they can each exist without the other—but the conscious mental state M, which they compose, is *ontologically one-sidedly dependent*[15] on its constituent parts—i.e., it is

[15] This notion of one-sided dependence is derived from the mereological theory articulated in Edmund Husserl's third Logical Investigation (Husserl 1970). For more on this notion and its significance in Husserl's ontology see Simons (1982, 1987), Smith and Mulligan (1982), and Thomasson (1999, Chap. 2).

necessary that the conscious mental state *M* cannot exist without its constituent parts and the special relation *R* that these parts *contingently* share.

A conscious mental state with the ontological structure of Kreigel's $SOMT_{10}$ is such that a certain kind of self-misrepresentation is rendered impossible; specifically, it is not possible for a self-representation to represent a first-order mental state that actually does not exist at all. For the existence of some other mental state is guaranteed by the ontological structure of the contexts in which self-representations are tokened. But there is nothing in this ontological structure to prevent a self-representation's misrepresenting some or even all of an outer-representation's phenomenal properties.[16]

Now, I argue, following Weisberg (2008, pp. 167–168), that even the possibility of non-empty or partial self-misrepresentation, like that which Kriegel's model allows, makes a self-representational model of phenomenal consciousness such that it still falls victim to certain debilitating problems. As Weisberg points out, this possibility poses a problem for the theory's compatibility with the traditional notion of phenomenal quality that is much like the incompatibility between the higher-order theory and the traditional notion of phenomenal quality discussed above. According to the traditional notion, the role of phenomenal quality in the 'division of labor' associated with producing phenomenally conscious experience is to determine *what* the experienced phenomenal content is.[17] But Kriegel's model of phenomenal consciousness does not accommodate this conception of the role of phenomenal quality in the production of phenomenally conscious experience. For the *what* of the phenomenal content experienced by the subject is determined (at least in the case of the misrepresented properties) by the content of the self-representation alone, and the 'represented' phenomenal qualities of the outer-representation are left idle.

Moreover, this possibility of self-misrepresentation is also inconsistent with Kreigel's basic explanatory strategy. Just as on the naturalistic higher-order representational approach, the basic idea of *naturalistic* self-representationalism is to provide a reductive but non-eliminative explanation of phenomenal consciousness in terms of self-representation relations. But here again we see the representational relation between the self-representation and the outer-representation left idle in the explanation of what determines the experienced phenomenal content.

Could this self-representational model be saved if we were to beef up the ontological constraints on conscious mental states so that whenever a mental state is incorporated into a complex whole with the structure of Kriegel's $SOMT_{10}$, then it is ipso facto phenomenally conscious? Such a view is defended by Robert Van Gulick (2001, 2004, 2006). According to Van Gulick's model,

> transforming a nonconscious mental state into a conscious one is a process of *recruiting* it into a *globally integrated complex* whose organization and

[16] This is acknowledged as a consequence of the theories articulated in Caston (2002) and Kriegel (2003a), and is left open as a possibility for all the self-representational models presented in Kriegel (2006, footnote 60).

[17] See Neander (1998).

> intentional content embodies a heightened degree of *reflexive self-awareness*. The meta-intentional content, on the HOGS model, is realized not by a distinct and separate external meta-state but rather by the organization of the complex global state itself, and the object state is a component of the global state into which it has been recruited. (Van Gulick 2006, p. 24 italics in original)

In other words, a mental state *M** is made conscious simply by being integrated into another mental state *M* that has the ontological structure outlined in Kriegel's $SOMT_{10}$. The primacy of the integration process in the production of phenomenal consciousness is the main point of contrast with the models presented by Kriegel (2003b, 2005, 2006) in which phenomenal consciousness is considered to be due primarily to self-representation and only secondarily to the ontological structure of the mental states in which self-representations are tokened.

What is not clear, however, is that Van Gulick's model actually helps provide any answers to the problems of self-misrepresentation. One could interpret Van Gulick's position as saying that phenomenal consciousness is the product of the integration process alone. But, on this interpretation, Van Gulick's model attempts to avoid the problems of self-misrepresentation by abandoning the basic explanatory strategy of explaining phenomenal consciousness in terms of representation. Now, of course, this explanatory strategy is not required for a self-representational theory; all a self-representational theory needs to acknowledge is that conscious mental states necessarily have self-directed representational contents. But, I argue, dropping this explanatory strategy also renders Van Gulick's model implausible. According to the higher-order global state model, the marks of a member of a globally integrated complex mental state are that it has a high degree of *global integration* or what Daniel Dennett calls 'cerebral celebrity'. The more globally integrated a state is, the more its contents are connected with the contents of other mental states, and so the more accessible these contents are to other mental subsystems. Thus higher degrees of global integration make mental states "more widely and powerfully influential within the minds in which they occur" and so "typically able to have more effect on the overall system's evolving mental state and the organism's behavior" (Van Gulick 2006, p. 24). However, it is not clear that a mental state with a high degree of integration into a globally integrated complex mental state thereby qualifies as a phenomenally conscious experience.

Weisberg asks us to consider a case of an unconscious state of jealousy—i.e., a state we are not aware of having at all, in any way—that causes us to act rudely:

> My jealousy in this case certainly appears globally accessible. It controls my moods, my other emotions, my judgements and my perceptions involving the target of my jealousy. It even affects physiological reactions like my temperature and my rate of heartbeat and respiration. The state is not only available to a wide range of systems and processes; it is actively accessed by many of them. Furthermore, I am clearly implicitly self-aware of the state, in Van Gulick's sense. My jealousy shapes my interactions with my social environment. It determines my judgements, my perceptions, my behavioral reactions, even my speech. And this in turn feeds back onto my emotional state, affecting its evolution. (Weisberg 2008, pp. 177–178)

It could be contested, of course, that we do in fact have *phenomenal* awareness of such a mental state, even though we do not have a cognitive awareness of it, to wit, that we experience the jealousy even though we do not recognize it as such. But this distinction between phenomenally experiencing a mental state and recognizing the state as such would still not help clear up the questions of whether, say, visual stimuli delivered under masked priming, subliminal visual and aural stimuli, or chronic pain (that a subject does not always report being aware of), which are all, it seems, globally accessible in Van Gulick's sense, are also phenomenally conscious experiences.[18]

In light of these problems, one might be tempted to interpret Van Gulick's position as one that employs the representational explanatory strategy, saying that phenomenal consciousness is not a product of the integration process alone, but rather a product of the reflexive self-awareness that is an essential part of a higher-order global state. On this view, there would be two kinds of reflexive self-awareness found in conscious mental states: those that play a role in the production of phenomenal consciousness and those that don't (they merely function as higher-order representations—be they conscious or non-conscious—of other mental states). This would help Van Gulick answer the question of which globally integrated mental states are conscious—the conscious mental states would be those that include a conscious-making type of reflexive self-awareness—but it would not help provide answers to the problems of self-misrepresentation.

To my knowledge, Van Gulick says *nothing* about the nature of reflexive self-awareness that would be of any help in this regard except that the ontological structure of conscious mental states rules out the possibility of a "problematic mismatch" between the self-representation and its object state (Van Gulick 2006, p. 38). But Van Gulick does not explain what the differences are between problematic and non-problematic mismatches, and he does not explain how self-representations can avoid partial self-misrepresentation. So, even though Van Gulick's model has the resources to avoid empty self-misrepresentation (in virtue of the fact that globally integrated complex states have the ontological structure articulated in Kriegel's $SOMT_{10}$), it seems at the very least to suffer from the same problems found in Kriegel's model; namely, the problems that arise from the possibility of partial self-misrepresentation. For, if Van Gulick's model can avoid these problems, he offers no explanation of how this is so.

Gennaro's wide-intrinsicality view adopts the same ontological view of conscious mental states as that found in Kriegel's $SOMT_{10}$,[19] but it attempts to avoid the problems of self-misrepresentation by stipulating that self-representation—what he calls an unconscious metapsychological thought—is infallible (Gennaro 2004, pp. 58–62, 2006, pp. 242–243). Now, setting aside the differences from Kriegel's and Van Gulick's views that arise as a result of Gennaro's insistence that the 'self'-directed representational state is a higher-order *thought*, it seems to

[18] Weisberg mentions these cases as well at Weisberg (2008, p. 178).

[19] This is contrary to what Kriegel says about Gennaro's ontology of conscious mental states at Kriegel (2006, p. 150). See Gennaro's description of his own view in response to Kreigel at Gennaro (2006, pp. 222, 238–240).

me that Gennaro's idea that self-representation is infallible is a step in the right direction. And Gennaro's insistence that self-representations are unproblematically infallible is also rightly adopted on the basis of a distinction between the infallibility of pre-reflective self-representations and the fallibility of reflective or introspective self-representations. Gennaro writes concerning this,

> it *is* possible to separate the higher-order (complex) conscious state from its target mental state in cases of introspection…. This is as it should be and does indeed allow for the possibility of error and misrepresentation. Thus, for example, I may mistakenly consciously think that I am angry when I am "really" jealous. The WIV [wide-intrinsicality view] properly accommodates the anti-Cartesian view that one can be mistaken about what state one is really in. However, this is very different from holding that the relationship between *M* and *M* within* an outer-directed CMS [conscious mental state] is similarly fallible. (Gennaro 2006, p. 242)

But these observations alone are deeply unsatisfactory as answers to the problems of self-misrepresentation raised above. For what is needed is an explanation of *why* self-representations are infallible that does not make self-representations radically different in kind from outer-directed representations. He does indicate that the "pinning" of higher-order or, as I would prefer it, self-representational content to the relevant world-directs state is carried out by means of a *demonstrative content* (Gennaro 2006, pp. 58–59). I agree with this and I will spell out a view based on this insight later in this essay. But what is still lacking here is the same as what is lacking in the other two models I've considered in this section*: a mind to presenting a self-representational model of consciousness that shows how the integration of the ontological structure of conscious mental states and the logical structure of self-representations provides answers to the entire range of problems of self-misrepresentation.* Any representational model of consciousness—be it higher-order or self-representational—that leaves this vital point out of account will not have the resources to provide adequate answers to the problems of self-misrepresentation.

4 The inadequacies of descriptive and Russellian self-representational models of representational content

The arguments in the previous section demonstrate that a self-representational theory that can adequately answer the problems of self-misrepresentation must be one that articulates a logical structure of self-representation and an ontological structure of conscious mental states, each of which works in concert with the other to bring about the required infallibility of self-representation. The required infallibility must meet two conditions: first it must preclude the possibility of both empty and partial self-misrepresentation; second, it must not be due to the logical structure of self-representation alone – there must be some significant role for the ontological structure of conscious mental acts to play in the production of infallibility.

The reasons behind the first condition are clear from the considerations of the previous section: if a theory acknowledges the possibility of even partial self-misrepresentation, this creates problems of consistency with the traditional notion of phenomenal qualities and problems carrying out the reductive explanatory strategy. The reasons for the second stem from the desire for the theory to be naturalizable. If the reductive explanatory strategy is to be successful from a naturalistic point of view, it must utilize representational structures that are of the same type as those that are thought to be possibly naturalizable themselves. And the most direct route to that goal is to draw on those representational structures that are already considered likely to submit to a naturalistic analysis.

All naturalistic models of representation advanced to date have sought to account for misrepresentation by casting the veridicality of an individual mental representation as a contingent property of mental representations.[20] Intentional representations have an essential normative property of truth-directedness and any acceptable account of representation will have to provide an explanation of this property. So since the relations between representations and their objects are only natural/causal relations on the naturalistic view, these relations must be contingent. So if the infallibility of self-representation is to be consistent with a naturalistic foundation, it must be case that self-misrepresentation is a logical possibility—to wit, it must be such that, in some possible world, the self-representation misrepresents the object mental state tokened along with it in the context of a complex conscious mental state—but, *in the context of this world,* misrepresentation is impossible. In other words, self-representations must be only contingently infallible, i.e., it must be *metaphysically possible*, but (in this world) *contingently impossible*, for a self-representation to misrepresent its object.[21]

So what kind of representation could do this job? In this section, I will consider and reject three kinds of logical structure that are found amongst referring representations: in particular, the logical structure of non-definite descriptions, definite descriptions, and Russellian direct representations. I will articulate and defend the adequacy of what I call neo-Fregean direct representation for our purposes here in the next section.

It seems obvious that non-definite descriptions could not possibly do what we need. For we need a kind of representation that veridically represents the

[20] See, for example, Dretske (1995, Chap. 1) and Millikan (1984, pp. 85–94), and Fodor's criticism of the "Crude Causal Theory" of mental representation at Fodor (1987, pp. 99–102).

[21] The distinction I draw here parallels Chalmers's distinction between logical and natural supervenience at Chalmers (1996, pp. 34–38). There he writes, "B-properties supervene *logically* on A-properties if no two *logically possible* situations are identical with respect to their A-properties but distinct with respect to their B-properties" (p. 35). In other words, if A-properties logically supervene on B-properties, then, in any logically possible (i.e., logically coherent) world, if B-properties exist, then A-properties exist. Contrast this with natural supervenience, which Chalmers defines as follows: "In general, B-properties supervene naturally on A-properties if any two *naturally possible* situations with the same A-properties have the same B-properties" (p. 36). A naturally possible situation is "one that could actually occur in nature, without violating any natural laws" (p. 36). With this distinction in hand, I can express my point in this paragraph by saying that the infallibility of self-representation supervenes *naturally*, *but not logically*, on the structure and context of self-representations. For only then can we accept the ostensibly paradoxical thesis that it is both (logically) possible for representations to misrepresent and (naturally) impossible for (a certain sort of) self-representations to misrepresent.

phenomenal qualities of a particular outer-representation or set of outer-representations tokened along with it in the context of a conscious mental state. Non-definite descriptions, however, represent indiscriminately in this regard. They represent objects by representing properties of objects. So any object that has the properties that satisfy the truth-conditions of the non-definite description is thereby represented by the non-definite description, regardless of the time or place at which the object exists (and, if you are inclined towards Meinongian ways of viewing the universe of objects, regardless of whether the object exists or not).

Definite descriptions, on the other hand, single out only one object, but they are still inadequate because they too represent their objects indiscriminately. Like non-definite descriptions, definite descriptions refer to their object or set of objects by means of representing properties of objects; but, unlike non-definite descriptions, they (when successful) refer to only *one* object of set of objects by representing a set of properties that belongs exclusively to that one object or set of objects. The inadequacy for our purposes, however, is that self-representations with definitely descriptive content do not have the resources to differentiate between the phenomenal qualities of outer-representations that *now* exist in *this very* conscious mental state and those that exist at other places and times (or, if one is so inclined, do not exist at all). For example, it is possible for a definitely descriptive self-representation M^* to refer to another mental state $M^{\diamond}$ by means of representing a set of properties that only $M^{\diamond}$ possesses while M^* is tokened in the context of a complex mental state along with a mental state M°, which is distinct from $M^{\diamond}$. Thus the same problem arises with definite descriptive content as was encountered with non-definite descriptive content; namely, definitely descriptive self-representations do not reliably represent the mental state or set of mental states tokened along with it in the context of a complex conscious mental state.

Another option for self-representation is the logical structure of directly referring representation. Unlike descriptive representations, direct representations do not refer to objects by means of representing properties of the objects; rather they refer directly to the objects themselves. There are two models of the logical structure of directly referring representations. According to one, the representational content of the directly referring representation is identical to the object it represents. This is what I call the *Russellian* model of directly referring representations. According to the other, the directly referring representational *content* and the *object* the representation refers to are distinct. This is what I call the *neo-Fregean* model.[22] The Russellian model has received the loin's share of support and critical attention in the recent literature; its two most prominent proponents are John Perry (1977) and David Kaplan (1989). The neo-Fregean model, however, has received comparatively little attention amongst analytic philosophers (in fact, many today do not seem to realize that there are two models of direct reference to choose from here), but it has had a host of capable supporters in the history of 20th century philosophy, the earliest of which are outside the canon of analytic philosophy: Husserl and

[22] For more on the history behind this contrast between Russellian and Fregean views of semantic content see Coffa (1991, pp. 79–82, 89–93).

Gurwitsch amongst others in the phenomenological tradition,[23] Evans (2003) and Smith (1989, Chap. 5) in the recent analytic literature.

My argument for the inadequacy of the Russellian model hinges on the requirement that the infallibility of self-representations have the specific modal status discussed above: that it be metaphysically possible but contingently impossible for self-representations to misrepresent their objects. Russellian direct self-representation is not able to procure this status due to the *identity* of its representational content and its object. Given the identity of content and object, a mismatch between the content and object of a direct representation is logically impossible. For if there is a self-representation with no object, by Leibniz's Law, the representation will also lack representational content. And if the object mental state of the self-representation has green phenomenal properties, by Leibniz's Law, the content of the self-representation must be of a mental state with green phenomenal properties. The logical structure of Russellian self-representation excludes the possibility of a self-representation that represents a phenomenally green mental state that actually doesn't exist or that is actually phenomenally red. Thus, if Russellian self-representation is infallible, it is metaphysically necessary that it represent veridically.

Now, even though the Russellian theorist could not account for the contingent infallibility of self-representation, he could still account for the contingency of the fact that self-representations have a particular content. He could say, for example, that a particular self-representational state M^* only contingently represents the set of phenomenal properties P belonging to an outer-representational state $M^{\diamond}$ in this world, for the counterpart of M^* at some other world represents another set of phenomenal properties or is empty. With this, we seem to have room to account for the intuition that our phenomenally conscious experiences are only contingently the way they are, for my counterpart in some other world may have only empty self-representations or his self-representations may represent outwardly directed mental states with altogether different phenomenal properties.

And perhaps this is all the contingency we need built into our theory. The decisive question here, however, is whether such an account of self-representation is amenable to a naturalistic analysis of representation in general. The basic idea of providing a naturalistic basis for representation is to tell a story of how representational relations are built on a metaphysical foundation of natural relations between the mind-brain and its surrounding world. Therefore, the only way that the Russellian model of self-representation could be considered unproblematically naturalizable is if one also accepts the controversial idea that the natural relations grounding representations in this world—such as, for example, causal relations—are not contingent, but rather *metaphysically necessary* relations.[24] This way the metaphysical necessity of the infallibility of self-representation on the Russellian model could be accounted for in terms of the metaphysical necessity of the natural relations that ground it.

The consistency of the neo-Fregean view of self-representation with naturalism, however, is achievable without recourse to this or any other more controversial

[23] See Smith (1982) for a presentation of Husserl's theory; and see Gurwitsch (1985b) for Gurwitsch's discussion.

[24] For a defense of such a view see Shoemaker (1998) and Zimmerman (2000).

theses, as I will demonstrate in the course of the next two sections. And so, without compelling reasons to accept the truth of necessitarianism about natural relations (or, at least, about the set of natural relations that ground mental representation), it seems that a demonstration of the compatibility of the neo-Fregean model of self-representation with the claim that self-representations are only contingently infallible would render the neo-Fregean model the more plausible of the two.

5 Indexical neo-Fregean self-representation

The basic idea of the neo-Fregean view of representation in general has two parts: *one* is that the content of representation (sense) is distinct from the object of representation (reference); the *other* is that representational content bestows a special normative status on representations that renders them either veridical or falsidical.

Now one may think that the non-identity of content and object itself renders neo-Fregeanism incompatible with the project of naturalizing mental representation. Such an attitude is supported by a suspicion that neo-Fregean contents can only be understood as 'spooky', third-realm, Platonic entities—entities that are indeed incompatible with an austere naturalistic metaphysics of mind. But such a view of content is not required, and other naturalistically compatible views of neo-Fregean intentional content have been given. For example, there is Searle's property dualistic view according to which intentionality is an irreducible property of brain states (Searle 1983, Chap. 10, 1994); and there is the dual-aspect monism of David W. Smith according to which intentional mental states and brain states are two aspects of the same substance (Smith 1995). According to each of these views, intentional mental representation is a natural phenomenon, but this natural phenomenon is not reducible to a causal relation of some sort. It is still natural; it is just not causal. And while intentional content is still understood as abstract or Platonic on these two views, the relation of abstract content to concrete intentional brain states is one of type to token, a kind of relation that is not in itself problematic from a naturalistic point of view.

Another source of opposition to neo-Fregeanism is the externalist view of mental content. And, indeed, there is a genuine incompatibility here. The externalist's project is to understand linguistic meaning and mental representation without recourse to abstract Fregean senses or 'internal' intentional contents of any kind; and there is a natural compatibility and consonance between the semantic project of externalism and the metaphysical project of naturalism. But, as the two views referenced in the previous paragraph indicate, it is not required that a naturalist also be an externalist. They show that there can be neo-Fregean naturalists, that one can coherently think that a natural mind can have intensional components. However, this lack of a logical incompatibility between naturalism and neo-Fregeanism does not mean that reconciling neo-Fregeanism to naturalism is unproblematic. The basic, as yet unresolved, obstacle to such a reconciliation is to provide an adequate naturalistic basis for the peculiar normativity of representational content.[25] And

[25] Again see the references in footnote 2 above.

insofar as this is still lacking, it will be at least unclear whether neo-Fregean intentionality is amenable to naturalism. But this is a problem facing naturalism as such, not just neo-Fregean naturalism. So it is not a problem that places *only* neo-Fregean naturalism on shaky ground.

Putting to a side for now these complexities that arise in bringing neo-Fregeanism and naturalism together, I will show that the neo-Fregean model of self-awareness has no obvious incompatibilities with naturalism, and that this gives *prima facie* reason for thinking that the neo-Fregean model of self-awareness is the appropriate model for use in a naturalistic self-representational theory of phenomenal consciousness.

According to the neo-Fregean account, the content of self-representation is distinct from the object of self-representation and the content of self-representation has a *two-dimensional* structure, much like that which Kaplan and Perry attribute to the logical structure of indexical utterances, except that the neo-Fregean rejects the identification of content and object that these two have built into their models. Thus it seems apt to say that neo-Fregean self-representation is a species of indexical representation.[26]

The two dimensions of indexical representational content are what Kaplan calls the *content* and *character* (Kaplan 1989) and what Perry calls *sense* and the *role* (Perry 1977) of indexical representation. To illustrate this distinction consider the indexical sentence "I am my father's first son". According to the Kaplan–Perry view, the *content* of this sentence is basically "what is said" or what has traditionally been called the proposition or sense expressed by the sentence (Perry 1977, pp. 475–477, Kaplan 1989, p. 500). Kaplan characterizes the meaning or content of a *sentence* as a function of the meaning of its parts (Kaplan 1989, p. 507). The meaning or content of the *parts* of an indexical sentence, on the other hand, are either the meaning of the logical constants, the meanings of the descriptive elements, or, where the sentential element is a directly referring expression, the object itself (*per* his Russellian view of representational content).[27] The neo-Fregean view of direct representation, on the other hand, departs from the Kaplan–Perry line insofar as the Kaplanesque content, meaning, or sense (as opposed to the character) is distinct from the object of representation. So the meaning or sense of the sentence "I am my father's first son" consists of a direct *representation* of the person uttering the sentence (not the speaker herself), the meaning of the logical connective (predication), and the descriptive content predicated of the speaker.

The character of an indexical representation (as opposed to the content or sense element), on the other hand, "provides a rule which determines the referent in terms of certain aspects of the context" (Kaplan 1989, p. 490); or, as Kaplan expresses it elsewhere, an indexical representation's character is a function from contexts of use to contents (Kaplan 1989, pp. 505–507). So, in the case of the sentence "I am my father's first son", the character of the indexical term 'I' can be formulated as:

[26] This thesis has also been articulated and defended at length in Smith (1989, 2005).

[27] Or, more strictly expressed, the content of a directly referring term is a function from circumstances of evaluation to extensions (Kaplan 1989, pp. 500–505).

'I' refers to its speaker or writer.[28]

These two aspects of an indexical representation's content are distinct, but they work in concert with one another. An indexical representation requires a character in order to attain a determinate meaning or sense. For an indexical representation only represents some object or state of affairs directly by means of a semantical dependence on the context in which it is tokened. And so, unlike some representations, whose reference does not change with context (like names), the meaning of two type-identical indexical representations can be different, due to differences in their respective contexts. For example, the meaning of the sentence, "I am blue", uttered by Jones, is different from the meaning of that same sentence, uttered by Smith. It is the same sentence (type) with the same character, but when uttered in different contexts, they have different meanings.

Now, after a token indexical utterance obtains a content—i.e., after it has been appropriately deployed in a context in such a way that the utterances's character determines a representational content—it maintains this content when evaluated in different *circumstances* or, as Kaplan expresses it, in different possible worlds or "actual and counterfactual situations with respect to which it is appropriate to ask for the extensions of a given well-formed expression" (Kaplan 1989, p. 502). This distinction is a handy one because it gives us a way to articulate the truth-conditions of an indexical representation. Given a certain indexical content x of the term 'I' in the sentence "I am my father's first son", we can determine the extension of this sentence in some other circumstance—say, a circumstance in which x has an older brother—and thereby determine whether the sentence's truth conditions are satisfied with respect to that circumstance. So, the extension of Jones's indexical utterance "I am my father's first son" is, let's say, a state of affairs that satisfies the sentence's representational content, rendering that sentence true—a state of affairs, in which the entity Jones is James's first son (James is Jones's father's first name). But in a world in which Jones has an older brother, Smith, the extension of the indexical sentence is rather a state of affairs that renders the indexical utterance false, a state of affairs that does not satisfy the sentence's truth-conditions. The *content* of the indexical term 'I' is the *same* in both circumstances, but, due to the differences between the two circumstances, the truth-value of the indexical utterance differs.

So evaluating the truth-value of a sentence across circumstances is a radically different procedure from uttering the same type of indexical expression in two different contexts. The former procedure presupposes sameness of content across circumstances, while the latter may result in different contents for each context. This is so because the character of indexical representations is *context*-sensitive, but *not* circumstance-sensitive (Kaplan 1989, p. 506).

Applying this elaborate semantical structure to self-representation, I think that the following is an adequate formulation of the character or role of self-representation that determines the representational content or, broadly speaking, the meaning of self-representational mental states:

[28] This formulation of the character of the indexical term 'I' is, of course, inadequate in many respects. For example, the word 'I', written on the note "I will return in 10 minutes" that is hung on my office door, could be removed by my colleague and placed on his own door in order to indicate that *he* (not I) will return in 10 min. But we need not preoccupy ourselves with these issues here.

(I) A self-representation M^* represents the phenomenal qualities of a mental state $M^{\diamond}$, which is co-tokened with M^* in a context of a complex mental state.

This character gives self-representation just the right kind of sensitivity to context required to avoid problems of self-misrepresentation. It excludes possibilities of empty self-representation, for the self-representational state's character is a function that determines a representational content *only* in the context of a complex state. And, if we employ the ontology articulated in Kriegel's SOMT_{10}, such states are necessarily composed of more than one component mental state. So, in cases where the self-representational state is tokened outside the context of a complex mental state (if such are possible), it will have no representational content to contribute to the subject's phenomenal experience. But, in cases where the self-representational state is tokened in the context of a complex mental state, it will represent the mental state or set of mental states tokened along with it.

This formulation of character of self-representation also shows how neo-Fregean self-representation excludes possibilities of partial self-misrepresentation without recourse to the necessary veridicality of the self-representation encountered in the Russellian model. The representational content of the self-representational state necessarily represent only that mental state or set of mental states that occur along with it in the context of a complex mental state. But the contexts in which self-representations have determinate representational content—i.e., complex mental states—are themselves only contingent existences. So self-representations are always true in a certain kind of context, instances of which just so happen to exist in this world, but which do not exist in every possible world.

Another way of expressing this idea is to say that self-representations, along with certain indexical utterances, are true a priori but not true necessarily. In the Kaplan–Perry semantical apparatus of indexical representations, necessary truth is distinct from a priori truth. This is so because this kind of indexical representation is true in every (relevant) *context* (i.e., every context in which they also have representational content), but is *not* true in every possible *circumstance*, to wit, under every possible interpretation of the representation's extension (Kaplan 1989, pp. 538–539). The sentence Kaplan likes to use to illustrate this point, "I am here now," is true in every context of utterance in this world, for its truth depends on the fact that the agent of the utterance utters that sentence at a certain time and place. This utterance's truth is determined a priori by the logical structure of its character in concert with the context of a world with the right sort of temporal structure, the right sort of natural laws, and the right sort of agents of utterance; namely, agents in worlds that, when engaged in making utterances such as "I am here now," are identical to the person uttering the sentence. Bringing these background conditions for the a priori truth of this sentence to light helps us to see how it is not necessarily true. For since there are worlds with radically different spatio-temporal structures or in which the physical laws governing human agents are sufficiently different, there are circumstances in which this indexical utterance occurs, but where its content is false, e.g., worlds in which agent of utterance and the person who utters the sentence are not identical.[29]

[29] See Smith (1989, pp. 217–218) for a thought experiment that exploits this possibility.

In such worlds, the representational content is identical to that which is true a priori in the actual world, but the temporal or spatial structure of the world in question does not provide for the truth of this utterance in every case.

Kaplan suggests that the mistake of thinking that a priori true indexical utterances are also necessarily true rests on the assimilation of the semantic role of context to that of circumstance (Kaplan 1989, p. 509). As explained above, the context of an utterance determines the utterance's content in accordance with its character. Circumstances, on the other hand, do not determine content. Rather, the content of an indexical representation is, as Kaplan says, "simply *independent* of the circumstance and is no more a function (constant or otherwise) of circumstance, than my action is a function of your desires when I decide to do it whether you like it or not" (Kaplan 1989, p. 497).

Thus, self-representation is infallible only *in a context*; specifically, only in the context of a complex mental state where it serves to found the existence of the complex along with the mental state or set of mental states it represents. So the content of self-representation is not necessarily true. For the a priori veridicality of self-representational content depends on the contingent existence of complex mental states. The existence of such mental states is a background condition for the truth of self-representations and it is a background condition that is not satisfied in every possible world or circumstance in which we can evaluate the truth or falsity of a given self-representation.

6 Complex mental states, neo-Fregean self-representation, and naturalism

The self-representational model of phenomenal consciousness that I have presented above can be summarized as follows:

> A mental state M is phenomenally conscious iff (i) M is a complex whole composed of and one-sidedly dependent on two or more (nonconscious) mental states, M^*, $M^{\diamond}$…and M°, (ii) M exists in the actual world or some other appropriately similar world, (iii) M^* is a neo Fregean indexical representation of another component mental state or set of component mental states of M, and (iv) M^*'s representational content is determined in accordance with the indexical character or role articulated in (I).

Is there anything in this theory that would give us cause to question its naturalistic credentials? It seems that the attribution of the ontological structure of a complex whole should not be a source of trouble in this regard. For there are a variety of natural phenomena that exhibit this structure. Biological entities across the board—molecules, organic collections of molecules, organs, and (as Aristotle realized) even living organisms—have the mereological structure of a whole that is dependent on the existence of its parts *and* the existence of certain relations amongst its parts. Moreover this sort of structure also seems to be present in numerous

non-biological physical phenomena like electrical circuits, storm systems, gravitational fields, and solar systems, just to name a few.[30]

There is also now a growing body of literature on the dynamical systems approach in cognitive neuroscience that supports the idea that the brain contains a variety of phenomena that have the structure of complex wholes (Yoshimi 2004). And some naturalistic self-representational theorists have argued that some of these complex mind/brain structures, such as the synchronous firing of neurons that many cognitive neuroscientists think holds the answer to the so-called "binding problem", play a central role in grounding the existence of phenomenally conscious experience.[31]

It also seems unlikely that any trouble should arise from the indexicality of self-representation or from the two-part semantical structure of indexical self-representation. Even though the metaphysics of mental representation is not Kaplan's project, he suggests, plausibly enough, that the character of indexical utterances is shaped by linguistic convention (Kaplan 1989, p. 505). Now, social convention is usually considered to be an unproblematic source of norms for the naturalistic philosopher, but it is not available at the level of self-representation for the simple fact that self-representation is a non-linguistic and, in many cases, 'pre-cognitive' phenomenon.[32] But I see no reason not to suppose that the character or role of self-representation is shaped by the same natural forces that shape the functions other neurological and biological processes.

The most likely source of trouble for the naturalistic credentials of the view I am defending here is the idea that self-representation is neo-Fregean. But, as I indicated above, neo-Fregeanism is not logically inconsistent with a naturalistic view of mental representation. All that is required to make a neo-Fregean view of content compatible with naturalism is to not limit the naturalizing strategy to one of reduction to causal relations. Now, this is, of course, no mean task, but it is not—at least, according to what we know now—a logically impossible task. There are a variety of natural phenomena to serve as a basis for intentionality that are not just simple causal relations, and a few different naturalistic neo-Fregean views of content have been developed on these grounds.[33] So, it seems, the only *real* source of trouble here is the lack of a successful story about how neo-Fregean

[30] Psychology, linguistics, and phenomenology are other fields where complex wholes have been found to be quite common objects of investigation. In the field of linguistics it held a central place in the development of Jakobson's phonology (Smith and Mulligan 1982, pp. 61–65). A very well known and often discussed example in the psychological literature is the duck-rabbit gestalt. Also see the interesting history of the debate amongst Gestalt psychologists about the nature of complex (what they called "non-summative") wholes in Smith and Mulligan (1982, pp. 65–81).

[31] See Kriegel (2003b, pp. 489–493, 2005, pp. 46–51) and Van Gulick (2001, 2004, 2006).

[32] I here leave open whether prereflective self-awareness also serves as a source of epistemically privileged access to our mental lives. I suspect that it does in a certain limited domain, but that its primary function is to serve as a ground for *introspective* judgment and intuition, a self-representational state that, when phenomenally conscious, presupposes the structure of prereflective self-awareness. Providing sufficient arguments for this point would require more space than I have available here. So I leave it for another occasion.

[33] Again see the broadly naturalistic theories of mental representation put forward in Searle (1983, Chap. 10, 1994) and Smith (1995).

representational contents—including their peculiar normative properties—are produced and determined by natural mechanisms. However, despite the continuing failure of naturalizing philosophers in this regard, it seems to me that there is still room to rationally hope for the emergence of a philosophical Prometheus who finally sheds light on the naturalistic ground of intentionality. But, more to the point here, insofar as this lack is a problem for a naturalistic neo-Fregean theory of representation, it is a problem for every other naturalistic theory of representation.

7 Conclusion

Over the past few decades there has been a proliferation of naturalistic representational theories of phenomenal consciousness of both the higher-order and self-representational varieties. The problems of higher-order misrepresentation and self-misrepresentation presented above can be seen as countervailing forces stemming the growing tide of naturalistic theories that may be deemed acceptable. If the arguments in the previous sections are sound, then it seems that there is really only *one* theory of phenomenal consciousness that has the resources to avoid these problems: the complex state Neo-Fregean view. I have argued that there is no reason to believe that this theory of phenomenal consciousness is inconsistent with a naturalistic view of the mental representation. And so, it seems that the neo-Fregean complex state theory also provides us a way to adhere to a broadly naturalistic metaphysics of consciousness. As such, it seems to me that there is no question about the truth of the claim that the complex state neo-Fregean theory of consciousness has the explanatory edge over both its higher-order and self-representational competition.

As I have shown, the problems with other self-representational views stem from two sources: either they acknowledge that self-misrepresentation is a real possibility and so suffer the problems of self-misrepresentation or they provide an inadequate explanation of how misrepresentation is excluded. The complex state neo-Fregean view, on the other hand, is designed to avoid these problems by providing an explanation of exactly how self-misrepresentation is infallible and how such infallibility is only a metaphysically contingent property of conscious mental states. Nothing in this project of constructing philosophical models that can overcome problems that threaten its predecessors, while maintaining the same basic principles of its predecessors, seems *ad hoc* or otherwise methodologically questionable to me. And so I take myself to have adjoined yet another important and perhaps conclusive link to the chain of insights stemming from the problems of high-order and self-misrepresentation: that the *only* acceptable self-representational model of phenomenally conscious mental states is that provided by the complex state neo-Fregean theory I articulate and defend in this essay.

Acknowledgments I am greatly indebted to numerous discussions of this material with David W. Smith and to comments on earlier drafts from Sven Bernecker and Martin Schwab. I would like to dedicate this article to the memory of my friend and teacher Denny Bradshaw. He was my first philosophical mentor and the first to introduce me to the philosophical study of consciousness.

References

Bilgrami, A., & Rovane, C. (2005). Mind, language, and the limits of inquiry. In J. McGilvray (Ed.), *The Cambridge companion to Chomsky* (pp. 181–203). Cambridge: Cambridge University Press.

Brentano, F. (1995). *Psychology from an empirical standpoint*. London: Routledge.

Byrne, A. (1997). Some like it hot: Consciousness and higher-order thoughts. *Philosophical Studies, 86*, 103–129.

Caston, V. (2002). Aristotle on consciousness. *Mind, 111*, 751–815.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Coffa, J. A. (1991). *The semantic tradition from Kant to Carnap*. Cambridge: Cambridge University Press.

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

Dretske, F. (2002). Conscious experience. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 422–435). New York: Oxford University Press.

Evans, G. (2003). Understanding demonstratives. In P. Yourgrau (Ed.), *Demonstratives* (pp. 71–96). Oxford: Oxford University Press.

Fodor, J. (1987). *Psychosomatics; the problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Gennaro, R. J. (2004). Higher-order thoughts, animal consciousness, and misrepresentation: A reply to Carruthers and Levine. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 45–66). Amsterdam: John Benjamins Publishers.

Gennaro, R. J. (2006). Between pure self-referentialism and the (extrinsic) hot theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 221–248). Cambridge, MA: MIT Press.

Gurwitsch, A. (1985a). *Marginal consciousness*. Athens, OH: Ohio University Press.

Gurwitsch, A. (1985b). Outlines of a theory of 'essentially occasional expressions. In L. Embree (Ed.), *Marginal consciousness* (pp. 65–82). Athens, OH: Ohio University Press.

Husserl, E. (1962). *Ideas: General introduction to pure phenomenology*. New York: Collier Books.

Husserl, E. (1970). *Logical investigations*. London: Routledge.

Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893–1917)*. Dordrecht: Kluwer Academic Publishers.

Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481–563). New York: Oxford University Press.

Kriegel, U. (2003a). Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy, 33*, 103–132.

Kriegel, U. (2003b). Consciousness, higher-order content, and the individuation of vehicles. *Synthese, 134*, 477–504.

Kriegel, U. (2005). Naturalizing subjective character. *Philosophy and Phenomenological Research, LXXI*, 23–57.

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 111–142). Cambridge MA: MIT Press.

Kriegel, U. (2009). Self-representationalism and phenomenology. *Philosophical Studies, 143*, 357–381.

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. New York: Oxford University Press.

Levine, J. (2006). Conscious awareness and (self-)representation. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 173–198). Cambridge, MA: MIT Press.

Loewer, B. (1997). A guide to naturalizing semantics. In B. Hale & C. Wright (Eds.), *A companion to the philosophy of language* (pp. 108–126). Oxford: Blackwell Publishers.

Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Noûs, 32*, 411–434.
Perry, J. (1977). Frege on demonstratives. *The Philosophical Review, 86*, 474–497.
Rosenthal, D. M. (1991). The independence of consciousness and sensory quality. *Philosophical Issues, I*, 15–36.
Rosenthal, D. M. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 406–421). New York: Oxford University Press.
Sartre, J.-P. (1956). *Being and nothingness: An essay on phenomenological ontology*. New York: Philosophical Library.
Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review, 117*, 245–274.
Seager, W. (1999). *Theories of consciousness*. London: Routledge.
Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Searle, J. R. (1994). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Shoemaker, S. (1998). Causal and metaphysical necessity. *Pacific Philosophical Quarterly, 79*, 59–77.
Simons, P. M. (1982). Three essays in formal ontology: Essay I. The formalization of Husserl's theory of wholes and parts. In B. Smith (Ed.), *Parts and moments: Studies in logic and formal ontology* (pp. 113–159). München: Philosophia Verlag.
Simons, P. M. (1987). *Parts: A study in ontology*. Oxford: Oxford University Press.
Smith, D. W. (1982). Husserl on demonstrative reference and perception. In H. L. Dreyfus (Ed.), *Husserl, intentionality, and cognitive science* (pp. 193–213). Cambridge: MIT Press.
Smith, D. W. (1986). The structure of (self)-consciousness. *Topoi, 5*, 149–165.
Smith, D. W. (1989). *The circle of acquaintance: Perception, consciousness, empathy*. Dordrecht: Kluwer Academic Publishers.
Smith, D. W. (1995). Mind and body. In B. Smith & D. W. Smith (Eds.), *The Cambridge companion to Husserl* (pp. 323–393). Cambridge: Cambridge University Press.
Smith, D. W. (1999). Intentionality naturalized? In J. Petitot, F. J. Varela, B. Parchoud, & J.-M. Roy (Eds.), *Naturalizing phenomenology* (pp. 83–110). Stanford: Stanford University Press.
Smith, D. W. (2004a). Return to consciousness. In *Mind world: Essays in phenomenology and ontology* (pp. 76–121). New York: Cambridge University Press.
Smith, D. W. (2004b). The Cogito circa A.D. 2000. In *mind world: Essays in phenomenology and ontology* (pp. 42–75). Cambridge: Cambridge University Press.
Smith, D. W. (2005). Consciousness with reflexive content. In D. W. Smith & A. L. Thomasson (Eds.), *Phenomenology and philosophy of mind* (pp. 93–114). Oxford: Oxford University Press.
Smith, B., & Mulligan, K. (1982). Pieces of a theory. In B. Smith (Ed.), *Parts and moments: Studies in logic and formal ontology* (pp. 15–109). München: Philosophia Verlag.
Strawson, G. (2005). Intentionality and experience: Terminological preliminaries. In D. W. Smith & A. Thomasson (Eds.), *Phenomenology and philosophy of mind* (pp. 41–66). Oxford: Oxford University Press.
Thomasson, A. L. (1999). *Fiction and metaphysics*. Cambridge: Cambridge University Press.
Thomasson, A. L. (2000). After Brentano: A one-level theory of consciousness. *European Journal of Philosophy, 8*, 190–209.
Thomasson, A. L. (2005). First-person knowledge in phenomenology. In D. W. Smith & A. Thomasson (Eds.), *Phenomenology and philosophy of mind*. Oxford: Oxford University Press.
Thomasson, A. L. (2006). Self-awareness and self-knowledge. *Psyche*, *12*, 2. Retrieved May 2006, from http://psyche.cs.monash.edu.au/symposia/kriegel/2Thomasson.pdf.
Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.
Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
Van Gulick, R. (2001). Inward and upward—reflection, introspection, and self-awareness. *Philosophical Topics, 28*, 275–305.
Van Gulick, R. (2004). Higher-order global states (hogs): An alternative higher-order model of consciousness. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 67–92). Amsterdam/Philadelphia: John Benjamins Publishing.
Van Gulick, R. (2006). Mirror mirror—is that all? In U. Kriegel & K. Williford (Eds.), *Self-representational theories of consciousness* (pp. 11–40). Cambridge, MA: MIT Press.

Weisberg, J. (2008). Same old, same old: The same-order representation theory of consciousness and the division of phenomenal labor. *Synthese, 160*, 161–181.
Yoshimi, J. (2004). Field theories of mind and brain. In L. Embree (Ed.), *Gurwitsch's relevancy for cognitive science* (pp. 111–130). Dordrecht: Springer.
Zahavi, D. (1998). Brentano and Husserl on self-awareness. *Etudes Phenomenologiques, 14*, 27–28.
Zahavi, D. (1999). *Self-awareness and alterity: A phenomenological investigation*. Evanston: Northwestern University Press.
Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. Cambridge, MA: MIT Press.
Zimmerman, D. (2000). Shoemaker's argument for his theory of properties. *Facta Philosophica, 2*, 271–290.