URIAH KRIEGEL

CONSCIOUSNESS, HIGHER-ORDER CONTENT, AND THE INDIVIDUATION OF VEHICLES

ABSTRACT. One of the distinctive properties of conscious states is the peculiar self-awareness implicit in them. Two rival accounts of this self-awareness are discussed. According to a Neo-Brentanian account, a mental state *M* is conscious iff *M* represents its very own occurrence. According to the Higher-Order Monitoring account, *M* is merely accompanied by a numerically distinct representation of its occurrence. According to both, then, *M* is conscious in virtue of figuring in a higher-order content. The disagreement is over the question whether the higher-order content is carried by *M* itself or by a different state. While the Neo-Brentanian theory is phenomenologically more attractive, it is often felt to be somewhat mysterious. It is argued (i) that the difference between the Neo-Brentanian and Higher-Order Monitoring theories is smaller and more empirical than may initially seem, and (ii) that the Neo-Brentanian theory can be readily demystified. These considerations make it *prima facie* preferable to the Higher-Order Monitoring theory.

INTRODUCTION: TWO RIVAL THEORIES OF CONSCIOUSNESS

Perhaps the earliest truly modern theory of consciousness is Franz Brentano's. According to Brentano, a mental state *M* is conscious when, and only when, it is partly about itself. If *x* has a conscious experience of a purple ball, *x*'s experience not only represents that a purple ball is present, but also that it, itself, is an experience of a purple ball. Brentano's thesis is that all and only conscious states are self-representational in this way.

The capacity for self-representation Brentano attributes to conscious states strikes many current-day philosophers and cognitive scientists as a tad mysterious. A more recent approach along similar lines is the Higher-Order Monitoring theory. According to the Higher-Order Monitoring (HOM) theory, a mental state is conscious when, and only when, there is *another* mental state which represents it. The experience of the purple ball is conscious because *x* has *another* mental state, whose content is that *x* is having an experience of a purple ball.

The two approaches are similar in that they construe consciousness in terms of self-awareness: in both theories, a mental state *M* being conscious depends on *M* being represented. They differ, however, in that while Brentano requires that *M* be represented by *itself*, the HOM theory allows,

Synthese **134:** 477–504, 2003.

and in fact insists, that it be represented by a numerically *different* mental state. That is, while both agree that M must be represented by a mental state M^*, Brentano claims that $M = M^*$, whereas the HOM theorist claims that $M \neq M^*$.

Each of the two theories has certain advantages over the other. These will be discussed in Sections 1–2 below. To be in a position to decide between the two, however, it is essential to get clear on what the empirical difference between them is supposed to come down to. I will discuss this question in Sections 3–4. My answer will be, in a nutshell, 'not much'. Given the dialectics, however, this answer will turn out (in Section 5) to give a certain edge to the Brentanian approach to consciousness.

1. A NEO-BRENTANIAN THEORY OF CONSCIOUSNESS

Conscious experiences have several distinctive properties. One is the qualitative character they exhibit. Another is the powerful impact they have on short-term memory. Yet another is the peculiar self-awareness they involve.

This self-awareness is peculiar in that it is not the *explicit* self-awareness of the sort x has when x is engaged in attentive introspection or careful self-scrutiny. Such explicit self-awareness, distinctive of the more reflective episodes of our mental life, is intriguing enough, but it does not characterize our conscious experience with any notable generality. It shows up in our stream of consciousness only periodically, and most of the time the stream flows without taking explicit notice of itself.

There is a different, dimmer sort of self-awareness that accompanies conscious experience *at all times*. Unlike explicit self-awareness, it does not require focusing one's attention on oneself and one's internal goings-on. Rather, it is permanently buzzing at the background of our conscious life. It is an *implicit* self-awareness whereby the subject is aware of herself as the *experience owner*. Here is how Alvin Goldman (Goldman 1970, p. 96; italics original) characterizes this phenomenon:[1]

> [Consider] the case of [consciously] thinking about x .... In the process of [consciously] thinking about x there is already an implicit awareness that one is thinking about x. There is no need for reflection here, for taking a step back from thinking about x in order to examine it .... When we are thinking about x, the mind is focused on x, not on our *thinking* of x. Nevertheless, the process of thinking about x carries with it a non-reflective self-awareness.

When one has a conscious experience of a purple ball, one is already implicitly aware that one is having a conscious experience of a purple ball. This implicit self-awareness is characteristic of all conscious experience as such.[2]

In the phenomenological tradition, the two forms of self-awareness – the explicit and the implicit – are distinguished in terms of the way the self is represented. In the reflective, introspective, explicit form, one is aware of oneself as *an object* among others. One may have greater concern for this object, but this object (the self) is observed as one more item *standing opposite* of the experiencing subject – a mere *Gegenstand*. Things are different with non-reflective, implicit self-awareness. Here one is aware of oneself as *the subject* of conscious experience. The self is represented as *the thing that does the experiencing*.[3] This self, considered only insofar as it is the thing that does the experiencing, cannot be introspected. Whenever you try to take a step back and observe it, it takes a step back with you, as it were. Trying to introspect it is like trying to hop on one's own shadow.[4]

A full-blown theory of consciousness would have to account for all the distinctive properties of conscious experiences. In some cases this is relatively easy. Thus, their powerful impact on short-term memory can most certainly be accounted for with one or another story in the genre of boxes-and-arrows functionalism. But in other cases, the task is notoriously difficult. Thus there is little agreement among philosophers and cognitive scientists on the proper account for the qualitative character and permanent implicit self-awareness exhibited by conscious experiences.

For Brentano, the latter is the key property of conscious experience. "In the same mental phenomenon in which the sound is present to our minds," he writes in his *Psychology from an Empirical Standpoint* (Brentano 1874, p. 127), "we simultaneously apprehend the mental phenomenon itself". His theory of consciousness therefore targets primarily the phenomenon of permanent implicit self-awareness. According to Brentano, this phenomenon is explained in terms of a special intentional structure exhibited by conscious states, namely, their being intentionally directed at themselves. Whatever else a conscious state represents, it always also represents itself. It is this self-representation that makes a mental state conscious. This notion is embedded within a theory of consciousness which features three central theses:

The Cartesian Coextension Thesis:

(CCT) All and *only* conscious states are mental states.

The Dualist Thesis:

(DT) All and only conscious states are non-physical states.

The Self-Representation Thesis:

(SRT) All and only conscious states are self-representational states.

The first two theses state background assumptions about the logical relationship between consciousness and the mind, on one hand, and between consciousness and matter, on the other; the third thesis purports to explain the nature of the self-awareness involved in all conscious states.

It is important to see that Brentano's three theses are largely independent of each other. In particular, there are no overt relations of entailment between SRT and either CCT or DT. (Amie Thomasson (2002) does a nice job of disentangling the various tenets of Brentano's theory and isolating SRT.) This means that one can hold SRT without committing oneself to the other elements of Brentano's theory. Indeed, one can hold SRT in conjunction with the following theses:

The No-Coextension Thesis:

(NCT) All, but *not only*, conscious states are mental states.

The Physicalist Thesis:

(PT) All conscious states *are* physical states.

I take it that a credible defense of Brentano's Self-Representation Thesis would do well to conjoin it with these alternative background assumptions. NCT has always been the working assumption of cognitive scientists and PT is the thesis, widely accepted among philosophers of mind, that token conscious states are identical to token physical states.[5] I call the conjunction of SRT, NCT, and PT the *Neo-Brentanian Theory of Consciousness*.[6]

The central thesis of both Brentano's original theory and the Neo-Brentanian theory is the thesis of self-representation. According to Brentano, every conscious state has a dual representational content. Its main content is the normal content commonly attributed to mental representations. But it also has a (rather peripheral) special representational content, namely, its own occurrence. Here is how Brentano (1874, pp. 153–154) puts it:[7]

> [Every conscious act] includes within it a consciousness of itself. Therefore, every [conscious] act, no matter how simple, has a double object, a primary and a secondary object. The simplest act, for example the act of hearing, has as its primary object the sound, and for its secondary object, itself, the mental phenomenon in which the sound is heard.

When *x* consciously hears a distant bagpipe, *x*'s auditory experience represents primarily the bagpipe sound and secondarily itself.[8]

At this stage it is important to get clear on a certain ambiguity in the expression 'self-representation'. To say that a mental state *M* is self-representational may mean either (i) that *M* represents *itself*, or (ii) that

M represents *the self*. The passage quoted above suggests a reading along the former lines. There are other passages, however, that suggest the latter reading. Thus, in an appendix to the *Psychology*, written four decades later, and which we can therefore take to express his considered view, Brentano (1874, pp. 276–277) writes that "the mentally active subject has himself as object of a secondary reference regardless of what else he refers to as his primary object".[9] Here it is the self that is represented secondarily by *M* – the self, specifically, *qua* 'the mentally active subject' (i.e., the thing that does the experiencing). Most probably, though, *M* represents *both* itself and the self. That is, what *M* represents (secondarily) is *its own occurrence within the self*. Whatever the subject's conscious experience is primarily directed at, it is also directed at the fact that the subject, herself, is having such an experience. Self-representational content is therefore *de se* content. If the primary content of *M* can be expressed more or less by 'This is a purple ball' or 'A purple ball is present', the secondary content can be expressed more or less by 'I myself am experiencing a purple ball'.[10] This aspect of Brentano's model answers to the fact that the self-awareness implicit in conscious experience is awareness of oneself, not only awareness of the state one is in.

The Brentanian model has its phenomenological attractions, then: it captures well the peculiar self-awareness characteristic of consciousness. It is perhaps for this reason that several authors have recently defended one or another version of the Neo-Brentanian theory – Carruthers (2000, chap. 9), Caston (2002), Gennaro (1996, chap. 2), Kriegel (2002), Thomasson (2000), Van Gulick (2001), and Zahavi (1999). Before this recent wave, Smith (1986) stands out as an early exponent of the Brentanian approach in the Anglo-American world.[11]

Despite being phenomenologically attractive, the Neo-Brentanian theory is deeply problematic. There is an intuition that the notion of self-representation is *mysterious*. The accusation of mysteriousness is often made out of hand, but perhaps we can explicate it as follows. It is unclear how the capacity for self-representation is supposed to be implemented in a natural, physical system. Indeed, there are good reasons to think that *no* physicalist account of mental representation could make sense of the notion of self-representation. This is not so much an argument against Brentano's own theory, since Brentano held that conscious states are *not* physical states, so he would have no problem admitting that self-representational states are non-physical. Rather, this is an argument against the Neo-Brentanian theory, which attempts to combine SRT with a naturalist and physicalist perspective on consciousness. Since the Neo-Brentanian theory holds that conscious states are both (i) physical states

and (ii) self-representational states, it must claim that there is a way for self-representational states to be physical, that is, that physical states can self-represent.

This claim is dubious, however. Consider our best physicalist, or naturalist, accounts of mental representation. According to these accounts, mental representation involves a natural relation between two physical states. At one end, there is a state of the subject's brain; at the other end, there is a state of the subject's physical environment. Thus, when *x* has a(n occurrent) mental representation of a purple ball, this mental representation consists in a natural relation holding between a certain pattern of neural activation, which is a state of *x*'s brain, and the presence of the purple ball, which is a state of *x*'s environment. The environmental state, *E*, is the representational *content* of the mental representation, whereas the brain state, *B*, is the *vehicle* that *carries* that content. Different physicalist accounts differ in how they choose to elucidate the natural relation between vehicle and content. The simplest version is a straightforward *causal* account, according to which *B* represents *E* just in case *E* causes *B*. A more sophisticated version is the *informational* account, developed by Dretske (1981). According to Dretske, *B* represents *E* just in case the occurrence of *B*-type states is nomically dependent on the occurrence of *E*-type states, where this nomic dependence is unpacked as follows: *B*-type states nomically depend on *E*-type states iff *E*-type states cause *B*-type states, and nothing else but *E*-type states causes *B*-type states, in all nomologically possible worlds. Yet another naturalist theory is the *teleological* account, developed most comprehensively by Millikan (1984), according to whom *B* represents *E* only if *B* has been selected by evolutionary or learning processes to covary with *E*.[12]

The trouble is that none of these accounts can accommodate self-representation. A Neo-Brentanian theorist must construe self-representation as involving a brain state *B* representing its very own occurrence (among other things). *B* does not necessarily have to represent itself *as* a brain state, or indeed *as* anything. It only has to represent itself. The problem, however, is that the various natural relations appealed to in naturalist accounts of mental representation are relations that cannot hold between a state and itself. For they are all *anti-reflexive* relations. Consider the causal account. According to this account, for *B* to self-represent, that is – for *B* to represent *B* – is for *B* to be caused by *B*. But no state can cause its own occurrence, no state can bring itself into existence. So within the framework of the causal account, self-representation is impossible. The other two accounts appeal to natural relations which, while going beyond simple causation, do involve causal relations as necessary components, so

the same problem is bound to arise within their frameworks as well. For instance, according to the informational account, for B to represent B is for B-type states to be caused by B-type states, and by nothing else, in all nomologically possible worlds. One nomologically possible world is the actual world, so this condition entails that B-type states be caused by B-type states, and by nothing else, in the actual world. And this entails that B-type states must be caused by B-type states, which is, again, impossible. Therefore self-representation is impossible within the framework of the informational account. And similarly for the teleological account: selection is a causal process.

The general argument proceeds as follows: (1) according to all naturalist accounts, mental representation implies a causal relation between the representing brain state and the represented environmental state; (2) the causal relation is anti-reflexive; therefore, (3) no brain state can bear the causal relation to itself; therefore, (4) no brain state can represent itself. Call this the *Argument from Physical Implausibility*. If sound, it would refute the Neo-Brentanian account of consciousness.

2. MODERN HIGHER ORDER MONITORING THEORIES OF CONSCIOUSNESS

The Neo-Brentanian theory is perhaps phenomenologically adequate, but it is physically implausible. This may be the reason twentieth-century philosophers, with their emphasis on naturalizability, have by and large shied away from the Brentanian approach. Many of them, however, have extracted from it a core they found to be sound, namely, the idea that consciousness is a *self-scanning device*. This is the basis of the Higher-Order Monitoring (HOM) theory of consciousness, as defended by Armstrong (1968, 1981), Dennett (1969), Lycan (1990, 1996, 2001), Rosenthal (1986, 1990, 2002), and others. What the HOM theorists reject is Brentano's specific model of how the self-scanning is done. Instead of positing mental states which effectively scan themselves, HOM theorists have opted for a division of the representational labor: some mental states represent the environment, others scan those first-order states. When a mental state M is scanned, or monitored, or represented by a second-order state M^*, M is conscious. M^* need not itself be a conscious state, although it may be – in case it is itself represented by a third-order state M^{**}.[13]

Not any representation of M will render M conscious, of course. M must be represented by M^* *in the appropriate way*. What the 'appropriate way' is, is something HOM theorists debate about. Two conditions are agreed on by all. First, M^* cannot be acquired through conscious inference.

Thus, if x has a desire to kill her father, and she infers that she has such a desire on credible evidence offered to her by her psychotherapist, *x*'s desire does not thereby become conscious. Second, M^* must be 'roughly contemporaneous' – to use a phrase of David Rosenthal's devising – with M: *x* must harbor her two states more or less at the same time. If on one Sunday afternoon *x* has the thought that it is a nice day, and then a year later comes to harbor a second-order representation of that thought, the thought does not retrospectively become conscious. M and M^* need not be absolutely simultaneous (what is?), since M may play a role in bringing about M^*, but they have to be at least roughly contemporaneous. The exact extent of the 'roughly' will be eventually figured out by cognitive scientists.

In the HOM model, the self-awareness peculiar to consciousness is captured by the second-order mental representation. As with Brentano, to capture this self-awareness, one must construe its representational content as referring not only to the subject's mental state, but also to the subject's self. Rosenthal (1990, p. 471) is explicit on this:

> When a mental state is conscious, it is not simply that we are conscious of the state; we are conscious of being in that state. This places constraints on what the content of these Higher Order Thoughts must be; their content must be that one is, oneself, in that very mental state.

When *x* has a conscious experience of a bagpipe, the precise representational content of M^* is not simply that there is an experience of a bagpipe sound taking place. It is the *de se* content that one is, oneself, experiencing the bagpipe sound.

The division of representational labor in HOM theory makes it possible to rescue the notion that consciousness is a self-scanner without committing to the psychological reality of self-representational states. HOM theory can combine NCT and PT with the following thesis:

The Appropriate Monitoring Thesis:

(AMT) All and only conscious states are states appropriately monitored (i.e., states appropriately represented) by other states.

The mental representation involved in higher-order monitoring is of the same ilk as the one involved in the first-order representation of the environment. This means that HOM theory can incorporate any naturalist theory of mental representation in its account of consciousness. It can therefore boast the physical plausibility that the Neo-Brentanian theory cannot.

However, as happens so often in the theory of consciousness, the gain in physical plausibility comes with concordant loss in phenomenological

adequacy. Recall that the HOM theory is an attempt to account for the permanent implicit self-awareness involved in conscious states. But does it really account for this self-awareness? There are reasons to doubt this.

When M is appropriately represented by M^*, it is M^* that contributes the representation of the self, but it is M that is conscious. So M itself does not involve any representation of the self. There is no representation of the self in the conscious state. There is a representation of the self, but it is unconscious. What is represented unconsciously, however, does not show up in the phenomenology. Since the self-awareness we are interested in *does* show up in the phenomenology, it must be part of the subject's conscious state. The HOM theory thus fails to account for the self-awareness *implicit in* states of consciousness.

(The HOM theorist may respond by denying that there is any self-awareness which we experience in our consciousness. But this is to deny the reality of the very phenomenon we set out to explain, that is, the phenomenon of permanent implicit self-awareness. This eliminativist development of HOM theory is coherent, of course, but it cannot claim much by way of phenomenological adequacy.)

It is important not to misconstrue this piece of phenomenological evidence against HOM theory. What phenomenology reveals is not that the representation of the self is inherent in M. Rather, what phenomenology reveals is that the representation of the self is conscious. That is, it reveals that M^* is a conscious state. This still leaves open the possibility that M and M^* are two numerically distinct conscious states. But this suggestion suffers from serious problems. First and foremost, if M and M^* are conscious, there is a sense in which they are parts of x's overall state of consciousness. Second, If M is conscious in virtue of being represented by M^*, but M^* is itself conscious, then M^* must be conscious in virtue of being represented by an M^{**} and so on *ad infinitum*.[14]

This is connected to two closely related points. First, as I point out elsewhere (Kriegel 2002), given that M^* is ordinarily non-conscious, the fact that M^* represents the self does not amount to self-awareness at all, since self-awareness requires *conscious* representation of the self (that is, it requires that the representation of the self be part of a subject's conscious state). Second, as Goldman (1993) argues, any model that portrays consciousness as an extrinsic property of the conscious state is inadequate, because consciousness is intrinsic to the conscious state. According to the HOM theory, consciousness is an extrinsic property, a property conferred on it from without, from the second-order state representing it.[15] But if the self-awareness we are interested in is a part, or aspect, of the conscious state, it is likely to be an intrinsic property of it. A further phenomenolo-

gical drawback in the HOM theory, perhaps less distressing, is its excessive vulnerability to zombie objections, noted by both Goldman and myself (see also Rey 1988).[16]

This last point is perhaps not much to worry about, but the previous three are, to my mind, very damaging.[17] They cast HOM as a theory of consciousness whose physical plausibility is obtained basically by *ignoring* the phenomenon in need of explanation and the fundamental phenomenological facts about it. The four points discussed have a cumulative effect which bears heavily against the HOM theory. I call their accumulation the *Argument from Phenomenological Inadequacy*. This is not to say that objections to HOM theory cannot be mounted on grounds of physical implausibility (see Byrne 1996; Carruthers 2000, chap. 8). If I have focused on the phenomenological case against HOM theory, it is because it brings out more acutely the main trade-off between the Neo-Brentanian theory and the HOM theory. Yes, various holes, of varying sizes, can be poked in the physical plausibility of HOM theory or the phenomenological adequacy of the Neo-Brentanian theory. But at the end of the day, supporters of the latter lean on phenomenological adequacy and supporters of the former on physical plausibility.

3. HOW MUCH DIFFERENCE DOES THE DIFFERENCE MAKE?

Upon reflection, this dialectical situation is somewhat surprising. After all, the two theories are remarkably similar. Both are premised on the notion that for a mental state to be a state of consciousness is for it to be represented in a certain appropriate way. In both theories, the occurrence of a conscious state requires the occurrence of two representational contents, a first-order content and a higher-order content. Moreover, the theories can agree completely on what those contents are. When x has a conscious experience of a purple ball, both ascribe to x the first-order content expressed more or less by 'a purple ball is present' and the higher-order content expressed more or less by 'I myself am experiencing a purple ball'. In both theories, it is the higher-order content that is key to consciousness. The only disagreement is over the question whether the two contents are carried by one and the same vehicle or by two distinct vehicles.

In fact, the similarities go even further. To see why, consider the following objection to the Brentanian approach, due to Rosenthal (1993). According to Rosenthal, while the Brentanian model is at least *coherent* as long as we consider conscious perceptions and thoughts, it becomes downright incoherent when we consider conscious desires, fears, and other mental states involving a *non-assertoric attitude*.[18] Suppose x has a con-

scious desire to own a red convertible. Brentano's model says that *x* is in a state with the first-order content (expressed more or less by) 'to own a red convertible'[19] and the higher-order content 'I myself desire to own a red convertible'. However, given that *x*'s state is a desire, the fact that the higher-order content of the desire is 'I myself desire to own a red convertible' would mean not that *x* is *aware of herself* as desiring to own a red convertible, but that *x desires* herself to desire to own a red convertible. To have a desire whose content is *p* is tantamount to desiring *p*; so if *p* is the proposition that one has a desire to own a red convertible, having a desire whose content is *p* is tantamount to desiring that one have a desire to own a red convertible.[20] But this is doubly inadequate. First of all, *x* may have no desire to desire to own a red convertible, and may even desire *not* to desire to own a red convertible (she may consider it vain).[21] Second, whether or not *x* desires her desire, the model fails to deliver self-awareness. That a desire is being desired does not entail that it involves any sort of self-awareness.

The defender of the Neo-Brentanian model can be expected to respond as follows. When *x* has a conscious desire to own a red convertible, *x* does have a mental state with the contents 'to own a red convertible' and 'I myself desire to own a red convertible'. But *x*'s mental state is related differently to its two contents. It is related *desire-wise* to its first-order content and *assertorically* to its higher-order content. The notion of relation-to-content can be cashed out, as it ordinarily is, in terms of *directions of fit*. For a mental state to be desire-related to an environmental state *E* is for it to have a world-to-mind direction of fit to *E*, whereas for it to be assertorically related to *E* is for it to have a mind-to-world direction of fit to *E*.[22] The Neo-Brentanian's claim is that *x*'s conscious desire is a mental state with a world-to-mind direction of fit to the content 'owning a red convertible' and a mind-to-world direction of fit to the content 'I myself desire to own a red convertible'.[23]

As it happens, the existence of such mental states – with dual directions of fit – has already been admitted elsewhere in philosophy. In moral psychology, a number philosophers have attempted, following McDowell (1979), to combine an internalist view of moral reasons with a cognitivist view of them. According to internalism, a moral reason is an intrinsically motivating mental state, that is, a mental state with a world-to-mind direction of fit to the appropriate action. According to cognitivism, a moral reason is a cognitive state whose function is to cognize correctly a moral reality, that is, a state with mind-to-world direction of fit to moral reality. In holding both views, one commits oneself to the notion that moral reasons have a dual direction of fit. This is precisely what David McNaughton

(1988, p. 112) argues for (see also Dancy 1993). Now, this combination of internalism and cognitivism has been challenged, by Michael Smith (1987, p. 56), precisely on the grounds that it is incoherent to suppose that one and the same mental state may have two different and opposing directions of fit. However, Little (1997) has argued – persuasively, to my mind – that Smith's argument is unsound.[24] In any case, this debate is far from resolved, so Rosenthal's objection to the Neo-Brentanian theory is not decisive.

What this exchange brings out, though, is that the HOM and Neo-Brentanian theorists agree not only on the assignment of contents, but also on the assignment of attitudes. To retain the coherence of her account, the Neo-Brentanian must postulate that every conscious state involves *two* attitudes, one towards its first-order content and one towards its higher-order content. (In the case of assertoric conscious states, the two attitudes are of the same type, but they are nonetheless different tokens.) So the Neo-Brentanian must agree with the HOM theorist that the occurrence of a conscious state involves two distinct attitudes and two distinct contents. The remaining disagreement is becoming ever thinner, then: although there are two attitudes and two contents involved, the Neo-Brentanian insists that they are all anchored in a single vehicle.

At this stage, it may seem silly of the Neo-Brentanian to insist on there being only one vehicle involved. She may be accused of making a fetish of the notion that there must be only one mental state involved in consciousness, however many contents and attitudes that state sustains. But for the Neo-Brentanian, the single-vehicle thesis is not at all arbitrary: it is necessary in order to construe the self-awareness implicit in conscious states as internal to them, that is, in order to make this self-awareness implicit *in* conscious states. In fact, the astonishing similarity between the Neo-Brentanian and HOM models may be taken to provide an argument *for* the Neo-Brentanian theory, since the Neo-Brentanian theory, unlike the HOM theory, is phenomenologically adequate. If two theories of consciousness are similar in almost every respect, but one is phenomenologically adequate while the other is not, clearly the phenomenologically adequate one is preferable.

What are we to make of this dialectical standoff? One thing we could do is stop worrying about which of the two theories is preferable, citing their substantial agreement. This is unsatisfying, however, because despite the extensive agreement between the two theories, their source of appeal is very different. The Neo-Brentanian theory captures marvelously the phenomenological facts about consciousness, whereas the HOM theory holds the promise of an eventual naturalization, long awaited and often

doubted, of the phenomenon of consciousness. This rivalry in sources of appeal makes the choice between the two theories more pressing. In the remainder of this paper, I will seek the kind of considerations that may bear on this choice. Since the choice is between positing one vehicle and positing two vehicles, I am led to the murky zone of vehicle individuation.

4. THE INDIVIDUATION OF VEHICLES

In fact, when we look more closely at the issue of vehicle individuation, the dissimilarities between the Neo-Brentanian and HOM theories shrink even further.

At the end of the last section, we saw that the only remaining difference between the Neo-Brentanian and HOM theories is that the former posits one vehicle and the latter two. The question, then, is How are we to count vehicles? It might be thought that the answer is straightforward. Recall that within the physicalist framework, vehicles are construed as brain states. Therefore, vehicle individuation reduces to the individuation of brain states. And as regards the individuation of brain states, the following proposal naturally suggests itself: a brain harbors two brain states at a given time t if there are two spatially discontinuous areas of the brain engaged in unusually intense neural activity, that is, two spatially discontinuous brain areas in which neurons fire their electrical impulses at increased rates (significantly higher than the baseline rate). If there is only one brain area in which there is such increased firing rate at t, then the brain harbors only one brain state at t. Call this the *Simple Proposal* for vehicle individuation.

On the Simple Proposal, the difference between the Neo-Brentanian and HOM theories comes down to this. According to HOM theory, when x has a conscious experience of a purple ball, there are two 'roughly contemporaneous' neural events (of increased firing rates) N_1 and N_2 taking place in different (spatially discontinuous) parts of x's brain. N_1 bears the appropriate natural relation (causal, informational, and/or teleological) to the purple ball, while N_2 bears the same natural relation to N_1. According to the Neo-Brentanian theory, by contrast, there is only one neural event taking place in x's brain, which neural event bears the right natural relation both to the purple ball and to itself.

The Simple Proposal will not do, however. To see why, I now digress to discuss one of the hottest research areas in cognitive neuroscience, namely, the so-called *binding problem*. Different parts of the brain are specialized for detecting and representing different kinds of environmental features. Thus, shapes are detected and represented in one area of the brain, colors in

another, motion in a third area. Needless to say, all these areas are different from the areas specialized for detecting non-visual features, such as flavors and sounds. In fact, the neural architecture of the brain is so functionally specialized that different aspects of shape are detected and represented in different areas. The horizontal line of an angle is represented by one neural assembly, the vertical line of the same angle by another. In some cases (e.g., the two lines of the same angle), the different brain areas are spatially adjacent. But in other cases (e.g., shape detection versus motion detection), the relevant areas can be quite far apart. Suppose *x* experiences a purple ball rolling from left to right. The purple color of the ball is represented by a neural event in one part of *x*'s brain; the ball's spherical shape is represented by a neural event in another part of *x*'s brain; and the left-to-right movement of the ball by a third event in a third area. Each of these neural events consists in an increased rate of action potential firing (i.e., firing of the electrical impulse) in the relevant specialized part of the brain. Thus, the presence of *purple* is represented by an event in which certain neurons in area 17 – also known as V1 – of *x*'s brain increase the rate in which they fire their electrical impulse. But the left-to-right movement is represented by a similar event in what cognitive scientists call MT, or V5 (see Zeki et al. 1991).[25] Yet the purple color and the left-to-right movement, as well as the spherical shape, are all experienced as part of a single, cohesive state of affairs. They are experienced as the color, movement, and shape of *one and the same object*. Cognitive scientists face the problem of explaining how the brain binds together these three distinct bits of information, given that their processing takes place in different 'departments' of the brain; how it represents their 'togetherness' as aspects of one and the same object.[26] This is known as the binding problem.

One solution to the binding problem would hypothesize that there is a special area of the brain in which 'it all comes together'. The brain binds the three different information bits about the color, motion, and shape of the ball by sending them all to this special area. The problem with this solution is, first of all, that there seem to be no such area in the brain, and second, that a binding mechanism of this sort would be extremely cumbersome and inefficient.[27] A cleverer solution has been suggested by von der Malsburg (1981). The idea is that the brain binds the various information bits by synchronizing the action potential firing of neurons at different parts of the brain. The brain uses some sort of feedback mechanism to synchronize firings in different neural events when, and only when, these different neural events represent features of one and the same object.

Consider again *x*'s conscious experience of the purple ball rolling from left to right. This experience involves three neural events of increased

firing rates. Suppose, for instance, that the relevant neurons in *x*'s Area 17 and MT fire their electrical impulse at a rate of 40 Hz, i.e., every 25 milliseconds. Then since the firing of each electrical impulse takes (approximately) only 2 milliseconds, the neurons representing the ball's purple color (in Area 17) might fire their electrical impulse, say, in cycles that come 15 milliseconds later than the cycles of firing by the neurons representing the ball's left-to-right movement (in MT). What the synchronization mechanism does is that it brings the relevant neurons in Area 17 and MT to fire their electrical impulse around 1 millisecond before or after each other. And so the brain represents that it is the same object which is both purple and moving from left to right. When their impulse firings are thus synchronized, the neural events in Area 17 and MT form part of one and the same cerebral representation, even though they take place in different (spatially disjoint) parts of the brain.

This solution is clever, in that it endows the brain with a new medium of representation for coding environmental information. Different environmental features are represented by firing *rates*, but the 'togetherness' of groups of features – their belonging together as features of one and the same object – is represented by firing *synchrony*. Thus, in representing the ball, the brain deploys *two* different representational media, the medium of firing rates and the medium of temporal correlation. Moreover, evidence aplenty has accumulated since von der Malsburg suggested this mechanism, showing that synchronization with the relevant functional role is actually taking place in the mammal brain (for a recent survey of the evidence, see Engel et al. 1999).

A number of neuroscientists have argued that the phenomenon of binding is key to understanding consciousness. In a celebrated piece, Crick and Koch (1990) sketch out a model of binding in visual experience, and proceed to offer it as a theory of consciousness. Crick and Koch do not indicate, however, why they think that non-conscious vision (e.g., in subliminal visual perception, or in visual perception of habituated stimuli, or in blindsight, etc.) does *not* involve binding. On the face of it, there is no reason to suspect that such non-conscious vision represents unbound, fragmented groups of features (see Revonsuo 1999). A step in the right direction is taken by Singer (1994, p. 101), who hypothesizes that only bound representations are eligible to reach the threshold of conscious awareness. That is, as long as information represented in the brain is unbound, it necessarily remains non-conscious. Engel et al. (1999) offer a survey of our current knowledge about binding and conclude that while we now have evidence that binding is a *necessary condition* for consciousness, we still have no evidence that it is a *sufficient condition* as well. That is,

all conscious states have a well bound representational content, but not *only* conscious states do: some non-conscious states have bound contents too. So the phenomenon of binding, while central to consciousness, is not distinctive of it, as Crick and Koch seem to imply.

The connection between consciousness and the phenomenon of binding is not as straightforward as pretended by Crick and Koch, then. After all, one could conceivably build a simple artifact which would incorporate some sort of synchronization, but the artifact is unlikely to exhibit conscious awareness. But my present interest in the phenomenon of binding is independent of Crick and Koch's hypothesis. It has to do with the fact that the psychological reality of binding falsifies the Simple Proposal, since it makes it possible for spatially discontinuous neural events to form a single brain state (a single vehicle). As I said, the ball's color is represented in area 17 and its motion is represented in MT. Area 17 and MT are quite far apart from each other, but the neural events occurring in them are part of one and the same brain state – a spatially discontinuous brain state, then – which represents a left-to-right moving purple object. Clearly, *x*'s perceptual experience of the moving purple object is a single mental state deploying a single vehicle to carry a single content. But this single vehicle is nonetheless constituted by a multiplicity of (appropriately synchronized) neural events. The Simple Proposal for vehicle individuation is therefore inadequate.

The Simple Proposal can be easily amended, however. According to a Less Simple Proposal, two neural events of increased firing rate, taking place in spatially discontinuous brain areas, constitute *one* brain state just in case the relevant firing rates are synchronized by the binding mechanism; they constitute *two* brain states in case the firing rates are *not* synchronized.

By discarding the Simple Proposal and embracing the Less Simple Proposal, the Neo-Brentanian and HOM theorists can agree on a further point. They can agree that when *x* has a conscious auditory experience of a bagpipe, there are two roughly contemporaneous neural events taking place in *x*'s brain: (i) neural event N_1, which bears the right natural relation (causal, informational, and/or teleological) to the sound of the bagpipe and (ii) neural event N_2, which bears the same natural relation to N_1. N_1 involves increased firing rate in the primary auditory cortex (also known as A1), while N_2 involves such increased rate elsewhere, probably somewhere in the frontal cortex.[28] The Neo-Brentanian and HOM theorists can agree on all this.[29] The only disagreement is over the question whether N_1 and N_2 constitute *two* brain states or just *one*. According to HOM theory, N_1 and N_2 constitute two brain states, whereas according to the Neo-Brentanian

theory, they constitute only one brain state. The interesting thing is that *this question is purely empirical*. It boils down to whether the action potential firings in N_1 and N_2 are synchronized or not. If they are, then N_1 and N_2 are bound into a single brain state; if they are not, N_1 and N_2 constitute two separate brain states. Now, whether or not the action potential firings in N_1 and N_2 are synchronized is an empirical question if anything ever was. There is no philosophical answer to it.

5. VEHICLE INDIVIDUATION AND SELF-REPRESENTATION

The conclusion of the last section is that the only disagreement between the Neo-Brentanian and HOM models is over a purely empirical matter. The empirical question of binding is a question philosophers have nothing to contribute in answering. There is therefore very little, if anything, for Neo-Brentanian and HOM theorists to quarrel over *as philosophers*.

As I said at the end of Section 3, however, the ever narrowing difference between the Neo-Brentanian and HOM theories provides a powerful argument for the former, given its superiority with respect to phenomenological adequacy. In this final section, I want to reexamine the objection that the Neo-Brentanian account will necessarily resist naturalization, in light of the above discussion of binding and its import on vehicle individuation.

The objection is based on what I called the Argument from Physical Implausibility. The general argument, as constructed in Section 1, depends on two premises: (1) all naturalist accounts of mental representation imply a causal relation between the representing brain state and the represented environmental state; (2) the causal relation is anti-reflexive. From these two premises it follows that no brain state can self-represent. However, the discussion in the previous section casts doubt on the soundness of this argument. If the discussion was on the right track, then the argument is unsound, because premise (1) is false. It is false because what naturalist accounts of mental representation imply is only that there must be a causal relation between *part of* the representing brain state and *part (or aspect) of* the represented environmental state. If so, the anti-reflexivity of causality is no longer a barrier to self-representation, since a brain state can be said to represent itself if *one part* of it represents *another* part of it.

Consider, for instance, the following *possible* three-step process, leading to a conscious experience of the bagpipe. At t_1, the sound of a bagpipe – a distal stimulus – triggers a relevantly appropriate causal process culminating in the occurrence of neural event N_1 involving an increase in the rate in which a certain subpopulation of neurons in the primary auditory cortex, or A1, fire their electrical impulse. As a result, N_1 represents the bagpipe

sound (in a naturalistically kosher way). At t_2, N_1 in its turn triggers the appropriate causal process, eliciting the occurrence of N_2, a neural event involving an increase in the firing rate of certain neurons in the frontal cortex. Again, the result is that N_2 represents N_1 (in a naturalistically kosher way). At this stage, then, x harbors two distinct brain states carrying the right sort of first-order and second-order representational contents. (This is, in fact, what the HOM theory says is involved in conscious experience. This is why HOM theory is invulnerable to the Naturalist Argument: because all it posits is two naturalistically kosher representational states.) But suppose that at t_3, the binding mechanism synchronizes the impulse firing in N_1 and N_2. That is, suppose it synchronizes the impulse firing by the relevant neurons in A1 and the frontal cortex. Then at t_3, N_1 and N_2 are bound into a single cerebral representation B. B is a single neural state which is both (i) a representation of the bagpipe sound and (ii) a representation of (i). This dual representational structure means that B has self-representational content, as required by the Neo-Brentanian model. The interesting thing, though, is that B is naturalistically kosher, since it is built up in a naturalistically kosher way.[30] B is a self-representing physical state, then. If so, the Neo-Brentanian theory does not necessarily resist naturalization.[31]

One immediate objection to this Neo-Brentanian reply to the Argument from Physical Implausibility is that it is overly speculative: we have no evidence whatsoever that any of this is actually taking place in the brain. This is certainly a good point, especially given the Neo-Brentanian's appeal to highly specific assumptions about neural mechanisms. On the other hand, the reply, whether or not empirically plausible, establishes one important thing. It establishes that there is at least one coherent scenario in which self-representation occurs in a natural, physical system. This already shows that the thesis of self-representation (SRT) is consistent with a naturalist-physicalist outlook. This is important, because it establishes that the power of physical states to self-represent is not conceptually impossible in the way suggested by the Argument from Physical Implausibility. If so, the Neo-Brentanian theory is not inherently mysterious. Whether or not self-representational states actually come into existence through the binding of neural events which partly represent each other, the fact that there is a way for them to come into existence without involving a super-natural relationship of self-representation is significant enough.

Another way to put the point is this. The proposed model of physical self-representation has both an empirical facet and a non-empirical facet. The following claim is empirical:

(ES) Consciousness arises from the binding of first-order and second-order representations.

But ES entails the following claim, which is not empirical in any interesting sense:

(NS) Possibly, consciousness arises from the binding of first-order and second-order representations.

Where the modality of the 'possibly' is metaphysical. ES is a testable hypothesis. It predicts, for instance, that blindsight is a result of breakdown of the mechanisms responsible either (i) for the production of second-order representations, or (ii) for the binding of first-order and second-order representations. But NS is not similarly testable. It is thus a non-empirical claim. But NS constitutes, all by itself, a refutation (by counter-example) of the Argument from Physical Implausibility.

Moreover, note that our speculative model is not significantly more speculative than the HOM theory of consciousness. The first thing to observe is that this Neo-Brentanian model is in fact almost identical to the HOM model. According to the latter, M is a conscious state iff it is suitably represented by M^*. But of course, we have no empirical evidence whatsoever that whenever x is in a conscious state, there is a non-conscious state occurring in her as well. The Neo-Brentanian Speculation suggests that M and M^* must be bound in order for consciousness to emerge. And admittedly, we have no empirical evidence for this. But we have no more empirical evidence that M^* exists at all than that it is bound with M. We are all in the zone of wild speculation here.

Another possible objection to the Neo-Brentanian reply to the Argument from Physical Implausibility is that it fails to deliver self-awareness, because it only explains how a mental state can represent itself, not how it can represent *the self*. However, the model readily extends to accommodate representation of the self. N_2 can be construed as representing the information that x herself is in N_2, not just the information that N_2 is taking place. If so, B will represent both the information that the sound of the bagpipe is present and the information that x herself is experiencing the bagpipe sound.

The Neo-Brentanian reply to the Argument from Physical Implausibility can be rejected on more technical grounds. For instance, it can be

claimed that the suggested mechanism for physical self-representation is itself physically implausible. One way to make the argument is to point out that while the binding of contents in the model is straightforward, the binding of attitudes is not. Recall that the Brentanian model distinguishes between the attitude of a conscious state towards its first-order content and its attitude towards its higher-order content. A conscious desire to own a red convertible involves a non-assertoric attitude towards the content 'to own a red convertible' and an assertoric attitude towards the content 'I myself desire to own a red convertible'. The binding of different representations with assertoric attitude is widely accepted, but what about binding assertoric representations with non-assertoric representations?

The problem with this line of argument is that it requires us to speculate *a priori* about what might and might not be physically possible for a natural system, in a way that has proven unwise in the past. All the same, as it happens we do have evidence of binding mechanisms operative in the motor system (Engel et al. 1999, p. 134) and, more importantly for our present purposes, of binding occurring *between* the visual area and the motor area (Engel et al. 1999, p. 138). Synchronization has been discovered to occur in the execution of tasks of visuomotor coordination. That is, neural events in the motor area have been discovered – by Roelfsema et al. (1997) – to be bound with neural events in the visual area to form a single visuomotor representation. If we take representations in the motor system to involve non-assertoric attitude, and representations in the visual system to involve assertoric attitude, the binding of representations from both areas entails the possibility of a bound representation with more than one attitude. According to the Neo-Brentanian, a conscious intention, say, to move one's hand, involves (i) a neural event N_1 in the motor area, representing the hand's intended movement, (ii) a neural event N_2 in the frontal cortex, representing the occurrence of N_1, as well as (iii) the synchronization of firing in N_1 and N_2. As I said, all this is wildly speculative, but the point is that the Neo-Brentanian has the resources to counter-speculate.

A different line of argument would be that once the Neo-Brentanian model is naturalized, it loses its phenomenological adequacy. An appropriately bound group of neural events is only a lump of neurons vibrating in the dark, and does not make up consciousness. There is no way to see how the thus orchestrated neural activity can yield anything like a subjective conscious feel. However, it is surely unfair to criticize a model of consciousness, first for its disallowing naturalization, then for its allowing naturalization. More to the point, it is important to recognize that what gives the Brentanian approach its phenomenological adequacy is not the denial of naturalizability, but the way it captures the fundamental facts

about the phenomenon of permanent implicit self-awareness. So far as I can tell, it is not a fact revealed in phenomenology that this sort of self-awareness is not, and cannot be, realized in physical substrate. Phenomenology has nothing to say about the types of substrate that might realize consciousness. It only comments on what it is like for the subject to be conscious.

6. CONCLUSION

The Neo-Brentanian theory of consciousness retains its phenomenological adequacy even when it is clearly seen to be compatible with the possibility of an eventual naturalization of consciousness. It is this phenomenological adequacy that is missing from the HOM theory. Given how similar the two theories are, I see no reason to prefer HOM theory over the Neo-Brentanian theory. The two agree that when a subject is conscious, there is a complex intentional structure instantiated in her. This intentional structure involves, they agree, one first-order representational content to the effect that some environmental state is occurring and another, higher-order representational content to the effect that one is, oneself, in a state that represents the occurrence of the environmental state in question. Further, it is agreed that this intentional structure involves two attitudes, one towards the first-order content and one towards the higher-order content, and that the attitude towards the higher-order content is always and by necessity assertoric, whereas the attitude towards the first-order content is not. Further yet, it is agreed that the two contents are vehicled by two (or more) roughly contemporaneous neural events in the subject's brain. The only disagreement is on the question whether the several neural events constitute two different brain states or only one brain state. This question is purely empirical, I have argued, and depends on the neurophysiological facts about binding. Therefore, the difference between the HOM and Neo-Brentanian theories is purely empirical. There is no *conceptual* or *principled* advantage in the HOM theory, as is pretended by its proponents. On the contrary, the Neo-Brentanian theory has the advantage of being much better positioned to capture the distinctive self-awareness implicit in conscious states. On the basis of these considerations, I recommend the endorsement of the Neo-Brentanian theory.

One could reject the whole approach taken by both the Neo-Brentanian and HOM theories, of course. It is possible to argue, and certainly has been, that consciousness has nothing to do with self-awareness and is solely a matter of qualitative character (e.g., Block 1995). Or one could predict that no reductive explanation of consciousness will ever work out (Chalmers

1995). These worries drop beyond the scope of the present paper. In this paper, I have argued that to the extent that we construe consciousness as a self-scanner, involving essentially a peculiar kind of self-awareness, there is little to distinguish the two major models of consciousness, and what there is gives a certain advantage to the Neo-Brentanian model of consciousness.

ACKNOWLEDGEMENTS

I would like to thank Victor Caston, David Rosenthal, and especially Timothy Hanks for various conversations that made this paper significantly better than it would have otherwise been. I would also like to thank two referees for *Synthese*, whose comments and suggestions have likewise improved the paper.

NOTES

[1] What Goldman calls 'reflective self-awareness' is what I call 'explicit self-awareness', and what he calls 'non-reflective self-awareness' is what I call 'implicit self-awareness'. Van Gulick (1988) also argues for a distinction between 'self-consciousness and introspection', the difference being that the former need not be reportable or have sentential structure. Van Gulick writes (1995, p. 278): "The content of my present [conscious] visual experience is not merely *of this room*, it is *of myself as a unified personal agent being aware of this room* . . .".

[2] With regards to this sort of self-awareness, there are interesting questions aplenty: In which animals can we find it? Is it a conceptual or non-conceptual form of awareness? etc. I am not going to discuss these questions here.

[3] This phenomenological distinction between the contents of implicit and explicit self-awareness follows Kant's distinction between what he called the *empirical ego* and the *transcendental ego*. The empirical ego is the self as mental substance constituted through the subjective processes of synthesis (applied through the inner sense). The transcendental ego is the self as that 'something' that is transcendentally responsible for such constitution. The transcendental ego is always the subject of experience, whereas the empirical ego is an object of (introspective) experience.

[4] This phenomenological characterization of permanent implicit self-awareness is certainly controversial, given its Kantian air. But the psychological reality of the phenomenon of permanent implicit self-awareness does not depend on this characterization, and the characterization is quite useful at least as a ladder to be thrown once climbed.

[5] Although this has been rejected in modern times by several writers, most criticism of identity theories have focused on the version requiring identity of types of mental states to types of physical states. Thus, Kripke's (1980) well-known argument against the identity thesis is explicitly directed at the type identity version. Similarly, there is no straightforward way to construe Nagel's (1974) master argument as an argument against

token identity, and certainly Nagel himself does not indicate that this is how he means his argument.

[6] I call it 'Neo-Brentanian' because it is Brentanian in spirit without being Brentanian in background. The spirit is captured by SRT, the background is captured by the assumptions concerning the relation between consciousness and mind and between consciousness and matter.

[7] The view is first introduced in Section 7 of chapter II ("Inner Consciousness") in Book 2, which is entitled "A Presentation and the Presentation of that Presentation are Given in One and the Same Act". Also: "... [T]he presentation of the sound is connected with the presentation of the presentation of the sound in such a peculiarly intimate way that its very existence constitutes an intrinsic prerequisite for the existence of this presentation. This suggests that there is a special connection between the object of the inner presentation and the presentation itself, and that both belong to one and the same mental act .... In the same mental phenomenon in which the sound is present to our minds we simultaneously apprehend the mental phenomenon itself .... We can say that the sound is the *primary object of the act* of hearing, and that the act of hearing itself is the *secondary object*" (Brentano 1874, pp. 127–128; italics original).

[8] In this model, the fact that the self-representational content is merely secondary answers to the fact that the self-awareness involved in ordinary conscious states is only implicit. Thus in other passages Brentano (e.g., 1874, pp. 128; 275–256) calls the self-representation 'incidental'. Self-representation is incidental, in that the awareness of self is psychologically peripheral, and the focus of attention is on the external 'primary' object of the experience. For modern proponents of the Neo-Brentanian approach, the difference between primary and secondary representation can ultimately be accounted for in terms of an attention mechanism responsible for the distribution of the subject's attention resources across different aspects of the situation she is in.

[9] This elucidation of the nature of secondary content is offered only in an appendix to a second edition (from 1911) of Brentano's book, and does not show up in the original version. It is for this reason that I take it to be Brentano's considered view. Fortunately, the appendix is added to the English edition of the book (from 1973).

[10] Some believe that *de se* propositional content is irreducible to non-*de se* propositional content (e.g., Castañeda 1966, 1969). Others disagree (e.g., Lycan and Boer 1975). I do not wish to take a stand on this issue here. But I claim that however it turns out, Brentano would construe the secondary content of conscious state along those lines: if *de se* content is irreducible, then Brentano would take the secondary content of conscious states to be irreducible.

[11] In the phenomenological tradition, the Brentanian approach is more common; see especially Husserl (1928) and Sartre (1937). (This might also explain Smith's endorsement.) William James seems to express similar notions in writing "Whatever I may be thinking of, I am always at the same time more or less aware of myself, of my personal existence" (James 1892, p. 42).

[12] Millikan's own version of the teleological account appeals only to evolutionary processes. Dretske (1988) modifies his original informational account to incorporate a teleological component, but he prefers appealing to selection processes of discriminative learning. A teleological account appealing to both evolutionary and learning processes (and more) is defended by Dennett (1987).

[13] Some proponents of the Neo-Brentanian theory – Gennaro and Van Gulick, in particular – consider their theories to be HOM theories. This is because they take the Neo-Brentanian

theory to be a version of the HOM theory. The way I use these names, however, 'HOM theory' denotes the view that the representing and the represented are wholly distinct states. There is no particular advantage to this usage, and I have no objection to the practice of considering the Neo-Brentanian theory a version of HOM theory. But since in this paper I am exploring precisely the difference between the view that consciousness involves one state and the view that it involves two states, I am preserving the name 'HOM theory' for the latter.

[14] This infinite regress was discussed already by Aristotle (see Caston 2002) and again by Brentano himself.

[15] Rosenthal (1990, 1991) defends this claim. He would not put it this way today (Rosenthal, in conversation), but still, there is interest in considering Rosenthal's theory as developed in the early nineties, since it is most probably the best articulated and most comprehensive development of HOM theory (and perhaps any theory) of consciousness one can find *in print*.

[16] Zombie objections are of course a dime a dozen: for any theory of consciousness, someone can be found who will readily conceive a zombie who satisfies the theory without exemplifying consciousness. But a theory that anchors consciousness outside the conscious state makes it that much easier to conceive of the occurrence of the relevant external conditions in the absence of consciousness. Also, just for the record, let us note one more unintuitive element in the HOM theory: that it posits an unconscious second-order state with every conscious state it admits.

[17] I count the points made at the previous two paragraphs as three different points, but the concerns they raise are, to be sure, largely overlapping.

[18] A mental state representing a certain state of affairs *S* has assertoric attitude if it represents *S as obtaining*; it has non-assertoric attitude if it represents *S* otherwise, e.g., *as desirable*. The distinction between attitude and representational content is needed to account for the similarities and differences between, say, the following three mental states: (i) believing is will rain, (ii) believing Smith is taller than Jones, and (iii) desiring it will rain. There is a certain similarity between (i) and (ii) and a *different* similarity between (i) and (iii). The way to capture these two different similarities is to say that the latter is similarity in content whereas the former is similarity in attitude. (For more detail, see Searle 1983.)

[19] I will be using this sort of construction to express the content of desires. With reporting the contents of beliefs, there is wide agreement that one should use *that*-clauses. Such agreement does not exist, unfortunately, for the reporting of desire contents. Here I will use to-clauses to express desire contents, but nothing I will say will depend on this particular way of expressing them. The reader is free to plug in their own preferred construction.

[20] That is, it would be better expressed by 'to desire to own a red convertible'.

[21] The logical relationship between first-order and second-order desires is impressively studied in Frankfurt's (1971) seminal paper.

[22] The notion of direction of fit has been introduced, I believe, by Anscombe (1957), and developed by Searle (1983, chap. 1). To say that a mental state has a mind-to-world direction of fit is to say that for its satisfaction conditions to be fulfilled, the state has to fit the way the world is. To say that a mental state has a world-to-mind direction of fit is to say that for its satisfaction conditions to be fulfilled, the world has to fit the mental state. In the former case, what is required is that the mind will change so that it enters the right state. In the latter case, what is required is that the world change so that the state the mind is in will be satisfied.

[23] This account can be extended to other non-assertoric attitudes by reducing those attitudes to combinations of beliefs, desires, and/or qualitative states. Take, for instance, conscious anger. Plausibly, *x*'s being angry that *p* involves *x*'s believing that *p*, *x*'s desiring that not-*p*, and a certain qualitative character (see Gordon 1987, p. 53). If so, a conscious anger that *p* can be construed as a mental state with (i) a mind-to-world direction of fit to the content *p*; (ii) a world-to-mind direction of fit to the content not-*p*; (iii) the qualitative character characteristic of anger; and (iv) a mind-to-world direction of fit to the content that *x* is, herself, angry that *p*.

[24] Smith argues that the notion of a mental state which has both mind-to-world and world-to-mind directions of fit is incoherent, because an internal state with a mind-to-world direction of fit normally disappears as long as the state of affairs to which it is directed is not the case, whereas an internal state with a world-to-mind direction of fit normally persist as long as the state of affairs to which it is directed is not the case. Thus, it is constitutive of *x*'s belief that Jones is eating an apple that it normally vanishes when Jones does not eat an apple, but it is constitutive of *x*'s desire that Jones be eating an apple that the desire normally persists as long as Jones is not eating an apple; in both cases, the normal persistence of *x*'s state is part of the satisfaction conditions of the state, which satisfaction conditions capture the state's content. According to Little, Smith's argument would be sound if *x*'s state was supposed to fit in both directions *the same content*. But there is no reason to suppose anything like that. The same point would apply in the present context: *x*'s conscious desire is supposed to fit in one direction one content and fit in another direction another content.

[25] This is, of course, just an example, and if it turned out that the representation of color and movement direction are not as specialized as all that, I would have to look for a better example. But this example is actually empirically well founded. Let me stress that, in general, everything is much messier in the actual neurophysiology of the human brain than it is here presented to be.

[26] When such binding fails, the subject may bring together features that do not belong together in the same object. Treisman and Schmidt (1982) have documented cases in which a subject forms a representation of a single object with a certain shape and a certain color when in fact the relevant shape and color are featured by different objects.

[27] Suppose *x* is presented with a purple ball rolling from left to right and a yellow ball standing still. How would *x*'s brain distinguish the color and motion that belong to the one ball from the color and motion that belong to the second, if all these bits are sent into the same brain area?

[28] We can quite safely speculate that second-order representations take place in the frontal cortex, because the frontal cortex seems to be the site of all highly sophisticated cognitive processes (e.g., reasoning). I take it that forming second-order representations is one of the most sophisticated things our brain can do, so it is likely to happen in the frontal cortex.

[29] They can agree on all this, though, only if we extend the notion of binding from the level of objects to the level of states of affairs. So far we have discussed the integration of features into objects. This is what binding research focuses on. But there may also be a binding process involved in the integration of objects and features into states of affairs. (Thomas Metzinger (1995) discusses the many cognitive levels at which something like binding may take place.) It is such integration that is required to make sense of the binding of N_1 and N_2, since the information they represent is not about the same object. It is worth noting that this sort of information integration *must* take place somehow, but not necessarily through the same binding mechanism that is involved in feature integration.

[30] What makes it naturalistically kosher is that there is no component of the brain state that represents itself. Instead, there is one part of a single brain state representing another part of the very same brain state. In a sense, then, there is no self-representation here at all. But I call this self-representation mainly because it does involve the special sort of content attributed to conscious states by the Neo-Brentanian theory.

[31] It is important to realize that although I have illustrated the way a Neo-Brentanian theory might be consistent with naturalism given the synchrony model of binding, the consistency does not depend on this model. What it depends on is the psychological reality of *some* mechanism of binding. However binding works, we can imagine first- and second-order neural events being bound into a single brain state in that way.

REFERENCES

Anscombe, G. E. M.: 1957, *Intention*, Blackwell, Oxford.

Armstrong, D. M.: 1968, *A Materialist Theory of the Mind*, Humanities Press, New York.

Armstrong, D. M.: 1981, 'What is Consciousness?', *The Nature of Mind* 55–67; Reprinted in Block et al. (1997).

Block, N. J.: 1995, 'On a Confusion about the Function of Consciousness', *Behavioral and Brain Sciences* **18**, 227–247; Reprinted in Block et al. (1997).

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

Brentano, F.: 1874, In O. Kraus (ed.), *Psychology from an Empirical Standpoint*, English edition. L. L. McAlister, trans. A. C. Rancurello, D. B. Terrell, and L. L. McAlister, Routledge and Kegan Paul, London (1973).

Byrne, D.: 1996, 'Some Like it HOT: Consciousness and Higher Order Thoughts', *Philosophical Studies* **86**, 103–129.

Carruthers, P.: 2000, *Phenomenal Consciousness*, Cambridge University Press, Cambridge.

Castañeda, H.-N.: 1966, ' "He": A Study in the Logic of Self-Consciousness', *Ratio* **8**, 130–157.

Castañeda, H.-N.: 1969, 'On the Phenomeno-Logic of the I', *Proceedings of the 14th International Congress of Philosophy iii*; Reprinted in Q. Cassam (ed.), *Self-Knowledge*, Oxford University Press (1994).

Caston, V.: 2002, 'Aristotle on Consciousness', *Mind*, forthcoming.

Chalmers, D.: 1995, 'Facing Up to the Problem of Consciousness', *Journal of Consciousness Studies* **2**, 200–219.

Crick, F. and C. Koch: 1990, 'Towards a Neurobiological Theory of Consciousness', *Seminars in the Neurosciences* **2**, 263–275, Reprinted in Block et al. (1997).

Dancy, J.: 1993, *Moral Reasons*, Blackwell, Oxford.

Dennett, D. C.: 1969, *Consciousness and Content*, Routledge, London.

Dennett, D. C.: 1987, *The Intentional Stance*, MIT/Bradford, Cambridge, MA.

Dretske, F. I.: 1981, *Knowledge and the Flow of Information*, Clarendon, Oxford.

Dretske, F. I.: 1988, *Explaining Behavior*, MIT/Bradford, Cambridge, MA.

Engel, A. K., P. Fries, P. Konig, M. Brecht, and W. Singer 1999, 'Temporal Binding, Binocular Rivalry, and Consciousness', *Consciousness and Cognition* **8**, 128–151.

Frankfurt, H. G.: 1971, 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* **68**, 5–20.

Gennaro, R. J.: 1996, *Consciousness and Self-Consciousness*, John Benjamin Publishing Co., Philadelphia/Amsterdam.

Goldman, A.: 1970, *A Theory of Human Action*, Princeton University Press, Princeton, NJ.

Goldman, A.: 1993, 'Consciousness, Folk Psychology, and Cognitive Science', *Consciousness and Cognition* **2**, 364-383; Reprinted in Block et al. (1997).

Gordon, R. M.: 1987, *The Structure of Emotion*, MIT Press/Bradford Books, Cambridge, MA.

Husserl, E.: 1928, In M. Heidegger (ed.), *Phenomenology of Internal Time-Consciousness*, J. S. Churchill (trans.), Indiana University Press, Bloomington, IN (1964).

James, W.: 1892, In G. Allport (ed.), *Psychology: The Briefer Course*, Harper and Row, New York (1961).

Kriegel, U.: 2002, 'Consciousness, Permanent Self-Awareness, and Higher Order Monitoring', *Dialogue*, forthcoming.

Kripke, S.: 1980, 'The Identity Thesis', in his *Naming and Necessity*; Reprinted in Block et al. (1997).

Little, M. O.: 1997, 'Virtue as Knowledge: Objections from the Philosophy of Mind', *Nous* **31**, 59–79.

Lycan, W. G.: 1990, 'Consciousness as Internal Monitoring', *Philosophical Perspectives* **9**, 1–14; Reprinted in Block et al. (1997).

Lycan, W. G.: 1996, *Consciousness and Experience*, MIT Press, Cambridge, MA.

Lycan, W. G.: 2001, 'A Simple Argument for a Higher-Order Representation Theory of Consciousness', *Analysis* **61**, 3–4.

Lycan, G. W. and S. E. Boer: 1975, 'Knowing Who', *Philosophical Studies* **28**, 299–344.

McDowell, J.: 1979, 'Virtue and Reason', *Monist* **62**, 331–350.

McNaughton, D.: 1988, *Moral Vision*, Blackwell, Oxford.

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

Metzinger, T.: 1995, 'Faster than Thought: Holism, Homogeneity, and Temporal Coding', in T. Metzinger (ed.), *Conscious Experience*, Inprint Academic, Thorverton.

Millikan, R. G.: 1984, *Language, Thought, and Other Biological Categories*, MIT, Cambridge, MA.

Nagel, T.: 1974, 'What is it Like to Be a Bat', *Philosophical Review* **83**, 435–450; Reprinted in Block et al. (1997).

Revonsuo, A.: 1999, 'Binding and the Phenomenal Unity of Consciousness', in *Consciousness and Cognition* **8**, 173–185.

Rey, G.: 1988, 'A Question about Consciousness', in H. Otto and J. Tueidio (eds.), *Perspectives on Mind*, Kluwer Academic Publishers, Norwell; Reprinted in Block et al. (1997).

Roelfsema, P. R., A. K. Engel, P. Konig, and W. Singer: 1997, 'Visuomotor Integration is Associated with Zero Time-Lag Synchronization among Cortical Areas', *Nature* **385**, 157–161.

Rosenthal, D. M.: 1986, 'Two Concepts of Consciousness', *Philosophical Studies* **94**, 329–359; Reprinted in D. M. Rosenthal (ed.), *The Nature of Mind*, Oxford University Press, New York and Oxford (1991).

Rosenthal, D. M.: 1990, 'A Theory of Consciousness', ZiF Technical Report 40, Bielfield, Germany; Reprinted in Block et al. (1997).

Rosenthal, D. M.: 1991, 'The Independence of Consciousness and Sensory Quality', *Philosophical Issues* **1**, 15–36.

Rosenthal, D. M.: 1993, 'Thinking that One Thinks', in M. Davies and G. W. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*, Blackwell, Oxford.
Rosenthal, D. M.: 2002, *Mind and Consciousness*, Oxford University Press, Oxford, forthcoming.
Sartre, J.-P.: 1937, *The Transcendence of the Ego*, F. Williams and R. Kirkpatrick (trans.), Noonday Press, New York (1957).
Searle, J. R.: 1983, *Intentionality*, Cambridge University Press, Cambridge.
Singer, W.: 1994, 'The Organization of Sensory Motor Representations in the Neocortex: A Hypothesis Based on Temporal Coding', in C. Umilta and M. Moscovitch (eds.), *Attention and Performance XV: Conscious and Nonconscious Information Processing*, MIT Press, Cambridge, MA.
Smith, D. W.: 1986, 'The Structure of (Self-)Consciousness', *Topoi* **5**, 149–156.
Smith, M.: 1987, 'The Humean Theory of Motivation', *Mind* **96**, 36–61.
Thomasson, A. L.: 2000, 'After Brentano: A One-Level Theory of Consciousness', *European Journal of Philosophy* **8**, 190–209.
Treisman, A. M. and H. Schmidt: 1982, 'Illusory Conjunctions in the Perception of Objects', *Cognitive Psychology* **14**, 107–141.
Van Gulick, R.: 1988, 'A Functionalist Plea for Self-Consciousness', *Philosophical Review* **97**, 149–181.
Van Gulick, R.: 2001, 'Inward and Upward – Reflection, Introspection, and Self-Awareness', *Philosophical Topics* **28**, 275–305.
Zahavi, D.: 1999, *Self-Awareness and Alterity*, Northwestern University Press, Evanston, IL.
Zeki, S., J. D. G. Watson, C. J. Lueck, K. J. Friston, C. Kennard, and R. S. J. Frackowiak: 1991, 'A Direct Demonstration of Functional Specialization in Human Visual Cortex', *Journal of Neuroscience* **11**, 641–649.

U. Kriegel
Department of Philosophy
Brown University
Box 1918
Providence, RI 02912, USA
E-mail: theuriah@yahoo.com