

## In defense of self-representationalism: reply to critics

Uriah Kriegel

© Springer Science+Business Media B.V. 2011

There are very few similarities between constructing a philosophical theory and laminating a driver's license. One similarity concerns their susceptibility to imperfection. When laminating a driver's license, often a bubble of air forms seemingly spontaneously to ruin the lamination's handsome sweep. One can press on the bubble in an attempt to flatten it, but the bubble reappears elsewhere on the surface. Pressed again, it reappears in a third location. Constructing a philosophical theory often involves a similar quixotic gambit. Of all the explicit and implicit desiderata we bring to the exercise, one or two are bound to frustrate one's construction. One can typically make some moves to satisfy the recalcitrant desideratum, but then another desideratum is suddenly left unsatisfied. And the process can be repeated.

I wish to start this response to my excellent—incisive, thoughtful, and creative—critics with a confession: my self-representational theory of consciousness, developed most fully in *Subjective Consciousness* (SC), has its own air bubble or two. Nonetheless, in responding to the critics I will conduct myself as though the theory is flawless and irreproachable; as though there are fully satisfactory responses to every objection they raise; as though all the theoretical sensibilities they bring to the table can be spoken to by the theory; as though my overall credence distribution was not updated, if barely perceptibly, after virtually every paragraph of their papers.

My critics' critical approaches are interestingly different. Brie Gertler develops a single sustained line of criticism against one central plank of the theory; Berit Brogaard offers a battery of quicker objections, counter-examples, and expressions of discomfort targeting a variety of aspects of the theory; Robert Van Gulick pursues a middle course of sorts. My response will go from the concentrated to the

---

U. Kriegel (✉)  
Department of Philosophy, University of Arizona, Tucson, AZ 85721, USA  
e-mail: theuriah@gmail.com

inclusive, and will assume such familiarity with the material as can be afforded by reading the *précis* in this symposium.

Brie Gertler targets my claim that a mental state is phenomenally conscious only if the subject is aware of it. My reason for holding this, recall from the *précis*, is something like the following reasoning: if a mental state of mine is conscious, then it is (like something) *for me*; but for the state to be *for me*, I must be aware of it; so, if a mental state of mine is conscious, then I must be aware of it. Gertler rejects the second premise: she accepts as datum that conscious states are *for their subjects*, but maintains that this datum can be accommodated without positing an awareness of those states. Indeed, she offers an alternative thesis to secure the conceptual link between consciousness and awareness, and argues that this alternative thesis accommodates also the other considerations (phenomenological and psychological) that I adduce in support of the claim that we are always aware of our conscious states (if ever so dimly). After describing Gertler's alternative, I will argue that it faces certain difficulties that would have to be addressed before we can take it to be a fully *viable* alternative.

Suppose one has a bluish experience of a twenty euro bill. On the thesis I defend, this necessarily involves being aware of one's bluish experience. Gertler's alternative is the thesis that having a bluish experience involves being aware of the property of bluishness. One does not need, in addition, to be aware of this property's instantiation by one's experience. This alternative thesis accommodates the conceptual link between consciousness and awareness more economically, says Gertler: a conscious state need not be an *object* of awareness—it only needs to be a *state* of awareness. As noted, Gertler proceeds to argue that the alternative, if true, would also explain various other data I adduce in favor of my thesis: the phenomenology of introspecting, the memorability of conscious experiences, the phenomenon of primesight, and the relationship between phenomenal and access consciousness.

One problem with Gertler's alternative thesis is that, although it accommodates the conceptual link between consciousness and awareness, the conceptual datum I originally adduced was more specific: a mental state's being conscious requires that it be *for me* (and not merely *in me*). It is this for-me-ness that I claim a state could exhibit only if one is aware of it. Gertler must hold that this for-me-ness can be somehow recovered without ever invoking the subject's being aware of that state. It is not entirely clear to me how this can be done.

I suspect Gertler would retort that if too much is read into this notion of for-me-ness, so that it becomes implausible to hold that a state can be for me without my being aware of it, then she would simply reject the idea that a mental state must exhibit this for-me-ness to be conscious. It remains that two ideas are independently plausible, however: first, that a state's being phenomenally conscious requires that there be something it is like *for me* to have it, and secondly, that for-me-ness cannot be exhibited by a state one is completely unaware of. It is on the independent plausibility of these two ideas that I pin the case for the thesis that conscious states are necessarily states one is aware of.

A second problem with Gertler's alternative view goes deeper, however: a variation on the epistemic argument against higher-order theories that I present in

Ch. 4 of SC seems to apply to it just as much. Let us divide mental states into three groups: those that are both states and objects of awareness, those that are only states of awareness, and those that are neither states nor objects of awareness. (In a first instance, this is meant only as a *conceptual* distinction, without commitment on the actual existence of such states). Suppose you undergo a bluish experience. According to Gertler, this means that you are in a state of the second category—a state of awareness. But how do you know that you are in such a state, a state of awareness (second category) rather than a state of *unawareness* (third category)? It seems you have some kind of direct phenomenological evidence for this, that you know this immediately and non-inferentially. Perhaps Gertler can provide an explanation for your having this kind of knowledge, but an explanation needs to be offered. The view that your bluish experience actually belongs to the first category—it is both a state and an object of awareness—has a readymade explanation: you know you are in a state of awareness because you are aware of being in it. Thus there is something epistemologically stable about the idea that conscious states are both states and objects of awareness. By contrast, the view that they are states of awareness but not objects of awareness is *epistemologically unstable*.

A final difficulty I wish to point out concerns a potential *metaphysical instability* in Gertler's alternative. The instability I have in mind pertains to the exact nature of the entity one is aware of when one has a bluish experience. Gertler's official view is that the qualitative property of bluishness is that entity, but she allows that awareness of such a property might consist in awareness of an external object (or surface, volume, film, or so on) that instantiates it, of that object's very instantiating of that property, or just of the property being instantiated. The first question we must ask, however, is what relationship Gertler supposes between the qualitative property of bluishness—the property instantiated by experiences—and the color property of blueness—the property instantiated by external objects. There are different ways to go here, and Gertler appears to wish to stay neutral on which way is best, but my suspicion is that her view runs into trouble whichever way we go.

Suppose that Gertler takes bluishness and blueness to be the same property. Then she must explain how this is possible; in particular, how a bluish experience could instantiate the very same property a blue surface does. This seems impossible on virtually every metaphysics of color. After all, experiences do not exhibit the reflective and refractive properties that blue surfaces do; they are not disposed to elicit (in normal observers etc.) bluish experiences numerically distinct from them; they do not have the categorical basis of such a disposition; and so on and so forth. To my knowledge, there is no remotely plausible metaphysical account of color properties that makes it plausible that an experience could instantiate them.

Suppose, on the other hand, that bluishness and blueness are numerically distinct properties. The advantage here is compatibility with the commonsense notion that experiences and surfaces do not instantiate the same (blueness-related) property. This, however, rules out the possibility that awareness of bluishness consists in awareness of an external object or an external object's instantiating of bluishness (unless all experiences are illusory). Since Gertler wants to cordon off awareness of the experience that instantiates bluishness, and of the experience's instantiating of

bluishness, she must hold that the relevant awareness of bluishness consists in awareness of the property itself (without awareness of any instantiations or instantiators of the property). In fact, if we construe states as property instantiations, and thus phenomenal states as phenomenal property instantiations, then the whole difference between Gertler's thesis and mine comes down to this: on Gertler's view, the subject of a bluish experience must be aware of the property of bluishness, whereas on mine, the subject must be aware of the relevant instantiation of that property. One might even say that Gertler's view requires awareness of the universal while mine requires awareness of the trope. I hope it is clear that at one level this is not a tremendous difference. At the same time, it strikes me that the trope version is more plausible than the universal version, in two ways: first, the universal version seems phenomenologically unmotivated (nothing in concrete phenomenology suggests relations to abstracta); secondly, the universal version puts in jeopardy the otherwise plausible notion that the kind of awareness involved in having a bluish experience is a *perceptual* (or at least quasi-perceptual) awareness—for it is unclear that we can have (quasi-) perceptual awareness of universals (the way we can of tropes).

I conclude that Gertler's alternative attempt to capture the conceptual link between consciousness and awareness is both epistemologically and metaphysically unstable, and that further development would be required before we can take it seriously as a competitor to the thesis that having a conscious experience requires being aware of it. Still, I have nowhere shown that no such further development is possible, and it remains to be seen whether a full-dress alternative might ultimately be devised. I now turn to consider Robert Van Gulick's critical discussion of the book.

Van Gulick's criticisms are four. I will address only three of them, as one boils down to a clash of intuitions (as Van Gulick acknowledges). In saying this, I do not mean to suggest that the contrary intuition is unconvincing or insignificant; merely that I have nothing to say by way of favoring my own intuition beyond what is in the book (at the beginning of Ch. 4). Van Gulick's three other criticisms concern my epistemic argument for preferring self-representationalism to higher-order theory, the principle of mental state individuation that would be needed to support such preference, and the prospects for a self-representational *reduction* of phenomenal consciousness.

A crucial premise in the epistemic argument is that the only evidence we could have for the proposition that all conscious states are represented is direct phenomenological evidence. The case for this premise proceeds by elimination: I argue against appeal to *indirect* phenomenological evidence, conceptual evidence, empirical evidence, or philosophical reasoning as alternative sources of support. But Van Gulick thinks that there are versions of these alternatives that I have not neutralized as successfully as one might wish.

The first alternative Van Gulick sketches is a version of the philosophical reasoning option. He offers the following reasoning: all conscious states are states we are aware of; awareness requires representation; therefore, all conscious states are represented. In response, I should note that although I do not discuss this particular version of the philosophical-reasoning option in the book, the version I do discuss (on p. 121) is very similar, and the general lesson I draw from that

discussion seems to me to clearly apply. The general lesson is that if we support the proposition that all conscious states are represented on the basis of philosophical reasoning from more general principles, the ultimate evidence for the proposition is effectively the evidence for those principles. In this case, the general principle is that conscious states are states we are aware of. What is the evidence for this principle? My claim is that it is direct phenomenological evidence—essentially for the reasons discussed in that part of the book (Ch. 4, Sect. 3, sub-section on “the first premise”). It cannot be a stipulative or conceptual matter, it cannot be a matter of evidence obtained through scientific inquiry, and it cannot be due to indirect phenomenological evidence. (To repeat, the reasons for this can be appreciated from the relevant sub-section of the book).

This brings me to the second alternative source of evidence Van Gulick considers—indirect phenomenological evidence. In the book, I consider the possibility that the proposition that all conscious states are represented is supported by inductive inference from *introspected* conscious states, which are clearly all represented (namely, by the introspective state), to *unintrospected* conscious states (pp. 118–9). Against it, I argue that such an inference would be unjustified, since the sample is gravely biased—what makes a conscious state belong to the sample is precisely that it is represented in introspection. Van Gulick finds this unpersuasive, however, claiming that all scientific inferences must proceed from the observed to the unobserved, and here introspection is the relevant manner of observation. (Joshua Weisberg, in his review of the book in *Mind*, also defends the viability of the relevant sort of inference, though on no particular grounds). Yet the fundamental flaw in the inductive argument under consideration is not just that it proceeds from the observed to the unobserved, but that the property projected through it pertains precisely to *being observed*. Compare: one can justifiably infer from the fact that all observed swans are white that all swans are white, but one cannot justifiably infer from the fact all observed swans are observed that all swans are observed; from the fact that all observed swans are at an eyeshot from an observer that all swans are at an eyeshot from an observer; from the fact that all observed swans are perceptually represented that all swans are perceptually represented. The problem with the inference from the fact that all introspected conscious states are represented to the fact that all conscious states are represented is that it is structurally akin to these obviously fallacious inferences.

Van Gulick's second criticism to be discussed pertains to the individuation of mental states. Since higher-order theory accounts for consciousness in terms of two distinct mental states, one of which represents the other, whereas my self-representational theory posits a single state with two parts one of which represents the other, it may seem that the difference between the two approaches comes down to whether two items count as two distinct mental states or as a single mental state with two parts—a matter of individuation. At the very least, the viability of my self-representational theory requires the availability of a principle of individuation that would warrant treating the two items as parts of a single state rather than two distinct states.

Van Gulick does not deny the availability of such a principle; indeed he mentions that in previous study he has provided one himself. His only claim is that the right

kind of principle has not been offered in my book. Thus this is not intended so much as an argument against my self-representational theory as against its author. My first reaction is to avow being perfectly pleased with the outcome of this dialectic: the theory has one more resource—Van Gulick's manner of individuating mental states—at its disposal. In addition, however, I should mention that I intended Ch. 6 of Subjective Consciousness (SC) to address this issue: Sect. 3 involves an extended discussion of the mereology of mental states, and offers specific claims about when a number of mental states compose a further mental state. In retrospect, I should have been more explicit on the direct implication of those claims about composition for issues of individuation. (There is, of course, the widely accepted view that composition is identity, wherefore composition conditions provide identity conditions, but none of this is explicitly discussed in the chapter). In any case, what I had in mind was that mental states compose a further state when they form a mereological complex rather than a mereological sum. The distinction between sums and complexes is discussed and applied to the present issue in pp. 221–224.

A final remark to make on this is that the disagreement between higher-order and self-representational theories is in any case not *only* about individuation. Another difference concerns the question of whether the higher-order item—whether a distinct mental state or part of the same state as the lower-order item—is *conscious*. Higher-order theories maintain that it is not, self-representational theories maintain that it is. The case for the latter stance is what the aforementioned epistemic argument offers.

Van Gulick's final line of criticism targets the reductive pretensions of self-representationalism. One problem he raises concerns the conceivability of zombies with suitable self-representing states; this comes up again in Berit Brogaard's paper, so I will postpone discussion. The other problem Van Gulick raises concerns my account of qualitative character. On the view I defend in Ch. 3, a conscious experience's property of having a certain qualitative character is identical to its property of representing itself to represent the right response-dependent property. This account requires a characterization of the right response-dependent property, which in turn requires a characterization of the relevant response. Characterizing the response in terms of qualitative (or phenomenal) properties would make it non-reductive. Characterizing it in terms of functional role faces other difficulties. In the book, I settle on characterizing the response in neural terms. Further theoretical pressures lead me to characterize it as a *disjunctive* neural kind, and correspondingly the response-dependent property as a *conjunctive disposition* (for details, see Sects. 6–7 in Ch. 3). On this view, the relevant neural responses are extremely heterogeneous, and the only reason to bring them together under a single disjunctive kind is that this allows the account to generate the right results regarding what experience has which qualitative character. Van Gulick is understandably dissatisfied with this way of proceeding, and claims that it involves implicit appeal to the qualitative or phenomenal properties of conscious experiences, which renders it non-reductive (see also Joseph Levine's review of the book in *Notre Dame Philosophical Reviews*).

My own dissatisfaction with the account of qualitative character I end up with is very much on the surface in SC, so I will not deny that there is something unlovely

about it. I do wish to deny that this renders the account non-reductive. For although one homes in on what disjuncts have to figure in the characterization of a qualitative character partly by considering the qualitative character itself, one can in the final count write down a list of all those disjuncts in purely neural vocabulary. The problem here is not quite that the account is non-reductive, then; rather it has to do with something like arbitrariness and/or informativeness.

This problem is nothing to sneeze at, of course, and to repeat I avow my displeasure fairly clearly in the book. The reason I adduce for embracing the resulting account nonetheless has to do with the unviability of *all* alternatives in logical space: characterizing the response in terms of a homogeneous neural kind (or for that matter functional role) returns the wrong results on what experience has which character; characterizing the response in qualitative or phenomenal terms forsakes the reductive hope altogether; characterizing the properties represented as response-*independent* rather than response-dependent is incompatible by the phenomenon of shifted spectrum (see Ch. 3, Sect. 5); characterizing qualitative character in altogether non-representational terms is incompatible with the transparency of experience (Sect. 2). Thus the only reductive account of qualitative character that returns the right results and is compatible with the transparency of experience and shifted spectra, I argue, is the one I offer. Its unloveliness counts against it, to be sure, but not as much as the other accounts' deficiencies do against them.

To conclude, Van Gulick's criticisms identify several pressure points for my self-representational theory, most notably the individuation of mental states and the characterization of internal responses in terms of which qualitative character is understood. Although I am inclined to think that the theory can withstand the pressure, this is certainly material worthy of future work.

Berit Brogaard's criticism is wide-ranging, but is organized under three central headings. The first targets my claim that the mystery of consciousness concerns primarily subjective character rather than qualitative character. My reason for claiming this, recall, is this: whereas qualitative character captures the identity condition of an experience (what makes it the experience it is), subjective character captures its existence condition (what makes it an experience at all); the mystery of consciousness pertains in the first instance to its existence rather than individuation; so the mystery pertains in the first instance to subjective character. Brogaard does not reject any particular step in this reasoning, but does find the result unacceptable. After all, it is completely mystifying that an experience's bluish qualitative character could result (whether causally or constitutively) from a bunch of neurons vibrating in the darkness of the skull.

As Brogaard mentions, in the book I claim that such explanatory gap as may concern neural activity and the bluishness of an experience attaches equally to quantum activity and the blueness of external objects. But she is unimpressed: perhaps the explanatory gap is no different, but then we have on our hands *two* mysteries rather than *zero*. In response, I must say that I am open to this possibility. My thought, however, is that this suggests that the mystery surrounding bluishness is not quite the mystery of *consciousness*—but some other kind of mystery. I suspect that the mystery of qualitative character is simply an aspect of the admittedly

mystifying question of how to reconcile the manifest image with the scientific image. This may well be a philosophical problem as deep as (or deeper than) the problem of consciousness, but it is a *different* problem: it is widely accepted that, of all elements in the manifest image, consciousness presents an additional, peculiar problem. It is only this additional problem that I contend pertains to subjective character rather than qualitative character. Thus to say that qualitative character does not generate the mystery of consciousness is not to imply that it does not generate *any* mystery.

Brogaard's second line of criticism focuses on the idea that consciousness comes in degrees. Brogaard claims that the self-representational theory of consciousness cannot accommodate this, because self-representation either occurs or does not occur—there are no degrees of being self-representing. My response is to divide and conquer: I wish to distinguish two senses in which consciousness may be said to come in degrees, and claim that the theory can accommodate one and need not accommodate the other. On a modest reading, the gradability claim means only this: once a mental state is conscious, its consciousness can be more or less phenomenally vivid, though there are no cases in which a mental state is not-quite-conscious but not-quite-unconscious. On a more radical reading, the state's very status as conscious can come in degrees, so that some states may be in some sense *sort of* conscious. To appreciate the distinction, consider two interpretations of a report of the form 'this state is vividly conscious and that one is mildly conscious': (i) 'this state is vividly-conscious and that one is mildly-conscious'; (ii) 'this state is-vividly conscious and that one is-mildly conscious.' In (i), the degrees qualify a state's consciousness, which it has fully. In (ii), the degrees qualify the state's very *having* of consciousness. The former is modest, the latter radical.

My contention is that the self-representational theory can accommodate the modest reading and should not accommodate the radical reading. It should not because it is implausible that a state's very status as conscious could come in degrees. Consciousness strikes me as akin to a light with a dimmer switch: one can turn the light up or down, and one can also turn the light off entirely, but there is no vague area where the light is neither determinately on nor determinately off. So the theory should not allow for such a vague area. At the same time, the theory can and does allow for degrees of luminosity when the light is on. It does so by appeal to a traditional conception of attention as a resource that can be distributed among the various items a person may represent at a time. The more of this resource is dedicated to a conscious state's representation of itself, the more phenomenally vivid the state is, on my view (for details, see Ch. 5 and 7 in SC).

Brogaard's third target is the theory's status as a metaphysical rather than empirical theory of consciousness. The theory aims not merely to characterize consciousness, but in some sense to capture its essence (hence the talk of identity and existence conditions). Brogaard offers three putative counter-examples. First, one can readily conceive of zombies whose internal states do not self-represent; this is, recall, something Van Gulick pressed as well. Secondly, there are actual cases of abnormal subjects who appear to lack any focal/peripheral distinction of the sort I appeal to in characterizing our awareness of our conscious states. Thirdly, there are actual cases of normal subjects with suitably self-representing yet unconscious states.

With regard to the self-representing zombies, my response is simply to accept that this provides evidence against the theory but hope that the evidence for the theory outweighs this evidence. Some philosophers dismiss conceivability data as illegitimate, others take (some of) them to entail facts about possibility. My view is intermediate: I take them to be legitimate but defeasible evidence for possibility facts. My hope, then, is that although the conceivability of zombies who satisfy my self-representational theory counts as a strike against the theory, the overall case I have presented in favor of the theory outweighs this evidence against it. (In addition, I should mention that, as discussed in Ch. 8 of SC, there are also perfectly coherent non-reductive versions of self-representationalism; namely, ones where phenomenal consciousness supervenes on suitable self-representation with nomological instead of metaphysical necessity).

The same response cannot be made with respect to the second and third cases adduced by Brogaard, since actuality does *entail* possibility (it is not merely evidence for it). However, as regards the case of the abnormal subject without peripheral awareness, I deny that her existence is an embarrassment to the theory. For the theory claims that having a conscious experience involves being *peripherally* aware of it only for some experiences—unintrospected experiences. As I say in Ch. 5 of SC, *introspected* experiences involve the subject's *focal* awareness of them. I do not consider in the book subjects who are “a-peripherally” aware of all items in their field of consciousness, and am grateful to Brogaard for bringing to my attention their existence. What the theory should say about such subjects, clearly, is that they are a-peripherally aware of their conscious experiences. This is not an *ad hoc* distortion of the theory, since the theory allowed all along for both peripheral and non-peripheral awareness of concurrent conscious experiences.

Finally, Brogaard argues that a normal subject can have unconscious (presumably dispositional) beliefs with the content <this very belief is a belief>. Such beliefs would be self-representational, and moreover in the way I claim is sufficient for consciousness: they would represent themselves *non-derivatively*, *specifically*, and *essentially* (see Ch. 4 for clarification of these qualifications). My response to this objection is twofold. First, in various places (Ch. 1 of my book *The Sources of Intentionality*, for instance) I deny the existence of dispositional beliefs, on the grounds that dispositions to believe can perform all the requisite explanatory work more economically. (Dispositions to believe, in turn, do not represent at all, but are merely disposed to represent). This means that for this to be a counter-example, we would need to have *occurrent* beliefs whose content is <this very belief is a belief>. But it is unclear in what circumstances our sub-personal belief-forming mechanisms might produce such beliefs. I deny, then, that unconscious beliefs with such a content are actual. The objector could move to the claim that, even if not actual, such beliefs are possible. But my claim would be that they are not: any belief that carried the content <this very belief is a belief> non-derivatively, specifically, and essentially would necessarily be conscious. Here too, I will concede that the conceivability data are on my opponent's side, but would pin my hope on the non-conceivability data outweighing the conceivability data at the end of the day. (I do recognize that it is not a trivial matter to show that they in fact do).

I conclude that although my self-representational theory identifies the problem of consciousness correctly and is perfectly consistent with a plausible gradability picture of consciousness, it faces a number of apparently conceivable counter-examples. In addressing those, it would have to either deny that these counter-examples are *ideally*, *competently*, or *non-superficially* conceivable, or else argue that despite being suitably conceivable they are not *possible*. In many cases, the latter route seems more plausible; in practice, pursuing it would involve showing that the evidence against the theory provided by the relevant conceivability data is outweighed by other evidence.

I started this response with a promise to conduct myself as though the self-representational theory I develop in SC is so evidently flawless that none of the difficulties raised for it had any merit to them. This veneer of methodological isolationism is probably less compelling by the time we have considered the variety of additional pressures and desiderata contributed by Gertler, Van Gulick, and Brogaard! Their commentaries make me think that the self-representational theory is perhaps best cast as a research program: a general framework that attempts to portray the broad outlines of what must be involved in a mental state's being conscious, but many of whose details are still in need of being worked out. These pertain most critically, to my mind, to the case for the principle that conscious states are necessarily states we are aware of (see Gertler), the sort of account of qualitative character that could be incorporated into the framework (see Van Gulick), and the treatment of the variety of conceivable scenarios in tension with the theory (see Brogaard).