Two Notions of Mental Representation

Uriah Kriegel

Forthcoming in U. Kriegel (ed.), *Current Controversies in Philosophy of Mind*, Routledge

Introduction/Abstract

The main thesis of this paper is twofold. In the first half of the paper (§§1-3), I argue that there are two notions of mental representation, which I call *objective* and *subjective*. In the second part (§§4-8), I argue that this casts familiar tracking theories of mental representation as incomplete: while it is clear how they might account for objective representation, they at least require supplementation to account for subjective representation.

1. A Parable

There is a possible world where, just as I was born, a brain neuroanatomically and neurophysiologically indistinguishable from mine was placed in a vat and fed random sensory stimulations by a machine suitably hooked to the vat. In fact, there are many, many such worlds. In one of them, the influx of sensory stimulation happens to be indistinguishable from the one my actual brain has enjoyed. Consequently, let us suppose, it is impossible to rule out from the inside that I am in fact such an envatted brain: the envatted brain's stream of consciousness is subjectively indistinguishable from mine. Thus whenever the stimulating machine is in state S1, the envatted brain undergoes an experience subjectively indistinguishable from the experience I normally undergo when I see an apple; when the machine is in state S2, it undergoes an experience indistinguishable from mine when I see a banana; when the machine is in S3, it undergoes an experience like mine when seeing a cherry; and so on and so forth.

An interesting question concerns what the envatted brain's S1-caused apple-ish experience represents. There are two conflicting views on this. A traditional view is that it represents exactly what my subjectively indistinguishable experience does: an apple. But following Putnam's (1981 Ch. 1) ingenious discussion, many philosophers have come to hold that it represents S1, the machine's state responsible for the apple-ish stimulation. The idea is, very roughly, that since the condition of the external world that stably covaries with and/or causes the brain's apple-ish experiences is S1, S1 is what experiences of that type represent in the envatted brain.

Which view is right? Does the brain's experience represent an apple or S1? My starting point is that this is really quite a silly question. There are two notions of representation, one on which the envatted brain's experience represents an apple and one on which it represents S1. In other words, the term 'representation' is ambiguous, and expresses two different concepts or notions. Accordingly, the term can be used in two different ways, to mean two different things. Both, however, are legitimate uses of the term 'representation': on the one hand, there is certainly a sense in which the experience represents S1, namely, the sense that the experience tracks the presence of S1; on the other hand, there is also a sense in which the experience represents an apple, namely, the sense that what the experience presents to the subject is apple-ish.

Thus the true moral of the brain-in-vat thought-experiment, it seems to me, is that a distinction must be drawn between two notions of mental representation – two different senses in which a mental state may be said to represent. The experiment is needed because in ordinary circumstances it is hard to distinguish the two notions. In my own experience of the apple, for example, what the experience presents to me and what it tracks are the same: an apple. It is only in the fantastic circumstances of the thought-experiment that an experience can be envisaged which presents to the subject one thing but tracks another.

2. A Distinction

There are no good terminological options for drawing the distinction brought out intuitively in the brain-in-vat thought-experiment. I will use the labels 'the objective

notion of representation' and 'the subjective notion of representation,' or *objective representation* and *subjective representation* for short. These terms are in many ways sub-optimal, but they will have to do.[1]

How to characterize the distinction in a more theoretically involved manner is, in a way, what this paper is about. For this reason, it would be unwise to prejudge certain issues by building commitments into the *definitions* of objective and subjective representation. Admittedly, we do need some way to fix our ideas regarding these two notions. But we can do so without unnecessary commitments through ostension of paradigmatic instances. One way to think of this is as offering a(n initial) model of the two notions not as necessary-and-sufficient-conditions notions, but as *prototype* or *exemplar* notions. We designate prototypical or exemplary objective and subjective representations – e.g., the envatted brain's – and consider a given mental representation objective and/or subjective if it is sufficiently relevantly similar to the prototype or exemplar.[2]

The distinction between objective and subjective representation is closely related to other, more familiar ones. Distinctions between personal and sub-personal representations, narrow and wide representations, phenomenal and psychological representations, may all turn out to be co-extensive with the subjective/objective distinction.[3] Even if they do, however, this should not be, and is not here, taken to be *definitionally* true. It is not *definitional* of subjective representation that it is personal, narrow, and phenomenal – nor of objective representation that it is sub-personal, wide, and psychological. My view is that the distinction between objective and subjective representation is deeper than all these, and underlies them, but I will not argue for this here.

Instead of arguing that the distinction is deep, I now argue that it is *thorough*. I want to suggest a four-way conceptual separability of objective and subjective representation: (a) there are conceptually possible scenarios where representation varies in the objective sense but remains invariant in the subjective sense; (b) conceptually possible scenarios where representation varies in the subjective but not objective sense; (c) ones where representation occurs *at all* in the objective sense but not in the subjective

sense; and (d) ones where it occurs in the subjective but not objective sense.[4] This is what I mean by the distinction being 'thorough.'

The brain-in-vat thought-experiment exemplifies (a): in the objective sense, the envatted brain's experience represents something my experience does not, namely S1, but in the subjective sense, the two represent the same thing, namely an apple. It might be objected that being an apple is a natural kind property, whose underlying nature involves imperceptible biochemical features, so nothing makes it the case that the envatted brain's experience (subjectively) represents an *apple* – rather than, say, a twin-apple (i.e., a fruit superficially akin to apple but with a completely different underlying biochemical nature). In response, I would concede that it is probably more accurate to say that the envatted brain's experience represents an apple-looking thing, rather than an apple. It does not denote the *natural kind* property of being an apple, but rather the *manifest kind* property of being 'apple-y' (or perhaps that of playing the apple role).[5] However, the same is true of *my* apple-ish experiences: in truth they only represent things as being apple-y, not as being apples.[6] So it is still the case that my experience and my envatted duplicate's are representationally type-identical in the subjective sense but not in the objective sense.

The following inverted-spectrum thought-experiment exemplifies (b). Imagine a world just like ours except for the following detail: your – or your counterpart's – color spectrum is inverted. As a result, during snowstorms your counterpart has an experience as of a black soft substance falling from the sky, while on sunny days the s/he has an experience as of peacefully yellow skies. These experiences track the same surface features of objects, but present to your counterpart very different features. Thus they are representationally type-identical to your experiences in the objective sense but representationally type-different in the subjective sense.

There are familiar examples of (c) – representation in the objective sense in the absence of representation in the subjective sense. The number of rings on a tree trunk tracks the tree's age, but does not present the age *to* the tree; thermometers' internal states track the ambient temperature, but do not present the temperature *to* the thermometers; and so on.

There are no familiar examples of (d), but consider the following thought-experiment. We can envisage a world where the only concrete particular is a disembodied soul 'floating about' in otherwise empty space, undergoing a random string of conscious experiences. In fact, we can envisage many, many such worlds. In one of them, the space soul's sequence of experiences is subjectively indistinguishable from yours. In a sense, it is impossible to rule out from the inside that you *are* in fact that space soul.[7] The space soul's experiences present to it exactly what yours present to you, but unlike your experiences, the soul's do not track anything.[8] Therefore, the space soul's experiences represent in the subjective sense but not in the objective sense.

This thought-experiment is similar to the brain-in-vat one, in that it too dissociates what an experience presents to its subject from what it tracks in the environment. But while in the brain-in-vat scenario the experiences still track *something*, in the space-soul scenario there is nothing for them to track. In a way, while the brain-in-vat thought-experiment shows that the *identity conditions* of subjective representation are independent from those of objective representation, the space-soul thought-experiment shows that their *existence conditions* are too.

3. Thesis

There are, in fact, many different notions of representation. We say of a reflection in a mirror or a puddle that it is an *imagistic representation* of some object or surface; we say of a graph or diagram that it provides a *mathematical representation* of some pattern; we say of a rainbow metaphor that it constitutes a *literary representation* of hope; and so on and so forth. There may well be a feature common and peculiar to all these kinds of representation, in virtue of which they all deserve the appellation 'representation.' But there are deep differences among them as well.[9] Most importantly, the nature of the representation relation implicated in each is very different. It is not as though literary and mathematical representations bear the same representation relation to their subject matters, with the former qualifying as literary simply because their subject matter is literary and the latter as mathematical because their subject matter is mathematical.[10] Rather, the very representation relation they bear to their subject matters is different.

My contention is that even within the realm of *mental* representation, we can distinguish two different senses in which mental items may be said to represent. For what makes a mental state track what it does, and track at all, is very different from what makes it present to its subject what it does, and at all. The feature in virtue of which mental states represent in one sense is completely different from that in virtue of which they represent in the other – and this is not a matter of different subject matters, since the subject matter can and often is actually the same.

If this is right, then seeking 'a theory of mental representation,' in the sense of a unified framework that accounts for a single relation which determines what, and that, mental states represent, may be as misguided as seeking a unified theory of literary and mathematical representation. What invites this misguide, it seems to me, is the fact that in this case the two fundamentally different kinds of representation are exhibited by the same vehicles: mental states. Ultimately, however, each notion requires its own account.[11]

The problem is that the theories of mental representation we have pursued most vigorously over the past forty years – those that fall under the rubric of 'naturalistic semantics' or 'psycho-semantics' – seem geared to account for the objective notion of representation exclusively, disregarding the subjective notion. I will now illustrate this by going through some of the most prominent options (§4). I will then lay out three possible reactions (§5) and then consider each (§§6-8).

4. Familiar Theories of Mental Representation

Theories of mental representation familiar from the 'naturalizing intentionality research program' tend to fall into two groups: causal-covariational theories and teleological theories. In its simplest manifestation, the causal-covariational approach claims that a mental state M represents a property F just in case Fs cause Ms under the right conditions (Stampe 1977). This kind of causal relation is most natural to appeal to in accounting for the tracking of external conditions. Observe, however, that the envatted brain's apple experiences do not have apples as their causes *ever* – i.e., under *any* conditions. Yet in the subjective sense what they represent are apples. So this approach seems wrongheaded as

an account of subjective representation. Note well: my present point is not quite that the causal-covariational approach lacks in principle the resources to account for subjective representation; rather that it is not geared to doing so, and appears to target instead (and quite plausibly) objective representation.

The best-known version of the causal-covariational approach is probably Fodor's (1990) 'asymmetric dependence' account: a mental state M represents a property F iff (i) it is a law of nature that Fs cause Ms, (ii) some Fs actually cause Ms, and (iii) if any non-Fs cause Ms, the fact that they do is asymmetrically dependent upon the fact that Fs cause Ms.[12] Again, this may be a promising account of representation in the objective sense, but not so much in the subjective sense. At the very least, condition (ii) is flagrantly violated by the envatted brain's apple experiences: none of them are caused by apples.[13]

A different account in the same spirit is Dretske's (1981) early informational semantics. According to it, M represents F iff M is nomically dependent upon F, that is, iff the laws of nature are such that M is not tokened unless F is instantiated. This comes very close to the standard account of tracking in reliabilist theories of epistemic justification (indeed, see Dretske 1971), and is thus a good candidate for a theory of representation in the objective sense. But as a theory of representation in the subjective sense it seems utterly inadequate: the envatted brain has experiences as of apples even when the property of being an apple is not instantiated. So it is false that they are not tokened unless the properties they (subjectively) represent are instantiated.[14]

The other familiar approach to mental representation is so-called teleosemantics. The idea, roughly speaking, is that a mental state represents in virtue of conferring the right kind of adaptive or reproductive advantage on the subject, or more accurately, in virtue of its 'correspondence' with external conditions conferring this kind of advantage.[15] Here again, there may well be much to recommend evolutionarily grounded tracking as an account of objective representation, but the prospects for such an account of subjective representation are on the face of it bleak. After all, if the envatted brain reproduces at all, it is certainly not in virtue of any correspondence between its apple experiences and apples, since there is none.[16] Note well: here again, I am not presently concerned to argue that teleosemantics is in principle incapable of accounting for the

subjective notion of representation; merely that clearly it is designed to account for the objective notion.

Consider Dretske's (1988) mature theory, which augments his original informational theory with a teleological component. According to the augmented theory, M represents F not iff M nomically depends upon F, but iff M is *supposed* to nomically depend upon F, where this means that (i) there is a motor response R, such that M has been recruited (through a process of 'discrimination learning') to have its present tokens cause R, (ii) past tokens of M nomically depended upon F, and (iii) it is the case that (i) because it is the case that (ii). This fails to accommodate the envatted brain's apple experiences in several ways. First, the envatted brain does not *have* motor responses (at least if motor responses are construed as bodily states), though it may seem to itself to have them. Secondly, past tokens of the envatted brain's apple experiences did *not* nomically depend upon any past instances of applehood, since there were no such instances. Finally, condition (iii) could certainly not be met, since there is no *reason* why the envatted brain has present token apple experiences – it is a pure accident.[17]

Another version of teleosemantics is Millikan's (1984, 1993) 'biosemantics.' Millikan's account is quite complex and not easily summarizable in a single cognitively surveyable biconditional, but a central *necessary condition* in the account appears to be this: M represents F only if there is a system S, such that (i) S consumes present tokens of M, (ii) past tokens of M occurred mostly when instances of F occurred, and (iii) S can perform its biological proper function because (i) and (ii) are the case. This necessary condition is quite obviously not met by the envatted brain's apple experiences. For starters, it is unclear how we might attribute a biological proper function to the brain's consumer system, since the brain faces no selection pressures. But more obviously, the envatted brain's past token apple experiences did *not* occur mostly when instances of applehood did.

There are other versions of teleosemantics (McGinn 1989, Papineau 1993 Ch.3), but I will not consider them here; I suspect they succumb to similar considerations. I conclude that both causal-covariational and teleological approaches to mental representation are geared towards its objective notion. The teleo-informational materials they employ are perfectly suited to account for the tracking of external conditions. (For

this reason, I will often refer to them in what follows as 'tracking accounts.') But to the extent that mental states can sometimes present to their subjects something they do not track, it is not clear that these materials are well-suited to capture what mental states present to their subjects. To repeat, my present point is *not* that no broadly teleo-informational story could ever accommodate subjective representation. It is rather that the familiar theories in that genre are so unnatural as accounts of subjective representation that it is most reasonable, and most charitable, to interpret them as not even *concerned* with subjective representation. This is doubly plausible given how genuinely promising they look as accounts of *objective* representation.

5. Notions and Properties

Agreeing that the familiar theories of mental representation in the teleo-informational genre are geared towards objective representation, but wishing to maintain that they fully account for all the relevant phenomena without need of supplementation, one might exploit the gap between notions and properties. After all, so far I have only argued for a *conceptual* distinction between two kinds of *notion*. I have not argued for a '*real*' distinction between two kinds of *property*.

At the same time, there is a trivial sense in which a conceptual distinction creates a *prima facie* presumption in favor of a real distinction. Consider again the envatted brain's apple-ish S1-caused experience. Suppose that, closely studying the vat's wiring, one person claims that the brain's mental state represents S1, while another, less talented for vat engineering, claims that it represents S27. It is natural to say that the first interpreter's ascription is true whereas the second's is false.[18] Likewise, suppose that, upon studying the neural correlates of consciousness in my brain and the envatted duplicate's, one person concludes that the brain's experience represents an apple while another, less neurologically talented, claims that it represents an elephant. Again it is natural to endorse the first interpreter's ascription and not the second's. Thus we have here two different true ascriptions and two different false ones. However, if two *different* ascriptions are true, then there are two different kinds of representational properties picked out by those ascriptions, each a constituent of a different truthmaker.[19] In this

way, the conceptual distinction between notions creates a defeasible presumption in favor of a real distinction between properties.

Given a presumptive distinction between two putative properties, several views on the ultimate relation between them are possible. Three stand out: eliminativist, reductivist, and nonreductivist. On the eliminativist view, ultimately there is no such property as subjective representation – the presumption is simply false. The subjective concept of representation may be useful in some way, but nothing in reality matches it. Accordingly, subjective-representational ascriptions are forsooth never strictly true, except at most in a minimalist sense.[20] On the reductivist view, there *is* such a property as subjective representation, but ultimately it reduces to objective representation, or some complex objective-representational structure.[21] Accordingly, subjective representation is an 'ontological free lunch,' fully grounded in objective representation. On the nonreductivist view, subjective representation is a real and genuine ontological addition over and above objective representation. (This is not to say, of course, that it does not reduce to *physical* properties; merely that if it does, it is not by reducing first to objective representation.) The choice can be appreciated by considering the following inconsistent triad:

1) There exists subjective representation.
2) Subjective representation is something over and above objective representation.
3) A theory of objective representation accounts for all the phenomena of mental representation.

Each of these is individually attractive, but they cannot all be true. Accordingly, any stable position on the matter must reject one of them. The eliminativist rejects 1, the reductivist 2, and the nonreductivist 3. I will now consider each of these options, arguing that any plausible accommodation of subjective representation would involve *something* importantly *un*familiar.[22]

6. Eliminativism

Eliminative positions are typically motivated by considerations of explanatory dispensability. Thus our eliminativist may argue that subjective representation would not explain anything, so there is no need to posit it.[23]

This eliminativist tack may be resisted by denying the explanatory impotence of subjective representation.[24] But more deeply, argumentation from explanatory dispensability presupposes a description of what needs to be explained. And the very description of the explananda is typically ontologically committal. Thus *explanatory* dispensability can support eliminativism about *x* only when combined with *descriptive* dispensability – the claim that there is no need to invoke *x* in describing what needs explaining. Eliminating subjective representation would thus require such a claim. Yet it is unclear how we can *describe* the scenario presented in the opening parable without mentioning a kind of representation shared by subjective duplicates. What needs to be explained in that scenario is precisely how an envatted brain stimulated identically to me is experientially presented with an apple. The very description of the explanandum thus invokes the subjective notion of representation.

The eliminativist may suggest that the apparent need to cite subjective representation in describing the scenario is an illusion, perhaps even an illusion that can be predicted from within the tracking framework of familiar theories of representation According to Rupert (this volume), for example, when we have second-order internal states that track our first-order representations, they can track only the presence of the state doing the representing, not the entity being represented. Accordingly, the second-order representation provides no genuine insight into what is being represented by the first-order representation, hence provides no support for the description of the brain-in-vat scenario as involving a representation of an apple.

However, when we consider introspectively our conscious experiences, they often (indeed typically) present themselves to introspection as directed at something outside the mind. Arguably, this is precisely the lesson of the so-called transparency of experience (Harman 1990). This is significant, because when we conceive of the brain-in-vat scenario, we seem to be employing a sort of first-person imagination whereby we imagine the envatted brain's mental life 'from the inside.' We imagine *being* the envatted brain and introspecting our own experiences while envatted. So insofar as the brain's

experiences are imagined as subjectively indistinguishable from ours, and ours are typically felt to be intentionally directed, the brain's ought to be typically imagined as seeming intentionally directed. To that extent, the natural description of the scenario involves mention not only of the brain's internal states but also of these states' *representational properties*.

In light of this consideration, the eliminativist may wish to call into question the possibility of the scenario presented in the opening parable. In conceiving the scenario, we conceive of a mental state whose representational properties in one sense are different from its representational properties in another sense. It is to describe this conceived scenario that we introduce the notion of subjective representation. But the eliminativist about subjective representation can resist the introduction of the notion by claiming that although the scenario is conceivable, it is not genuinely possible – it is not *metaphysically* possible. The view may be that, as a matter of Kripkean a posteriori necessity, the envatted brain's S1-caused apple-ish experience represents only S1; in no sense does it represent an apple.

Regardless of whether conceivability is generally a good guide to metaphysical possibility, however, the present objection is completely implausible. For an identically stimulated envatted brain ought to be not only metaphysically but *nomologically* possible. Consider that conscious experiences are widely acknowledged to have *neural correlates*, and it is certainly possible, consistently with the actual laws of nature, to duplicate the neural correlates of my stream of consciousness by duplicating exactly (i) the neural state of my brain at the beginning of my biological life and (ii) the sensory inputs I have enjoyed since. Everything we know about brain function suggests that this is nomologically possible, and would result in duplication of my conscious experience itself. The quality of the link between conceivability and metaphysical possibility is irrelevant.[25]

Conceding that the scenario from the opening parable is metaphysically and even nomologically possible, the eliminativist might insist that the property it requires us to posit nevertheless fails to qualify as representational. She may concede that one could use the word 'representation' to designate whatever one wished, but insist that the phenomenon picked out by the subjective notion is not representation *in the*

philosophically relevant sense. On this version of eliminativism, the envatted brain's state does present an apple to it, but it does not *re*present an apple in any philosophically significant sense.

This smacks of a verbal issue, but let us set that aside. Whether this kind of objection is plausible depends ultimately on what we require from a property to qualify as representational 'in the philosophically relevant sense.' If there is any substantive answer to this question, I think it would have to appeal to the traditional idea of *intentionality* (Brentano 1874).[26] The thought would be that 'the philosophically relevant sense' of representation is that which involves the features definitive of intentionality. Two definitive features stand out: (i) the feature that underlies failure of truth-preserving existential generalization; (ii) the feature that underlies failure of truth-preserving substitution of co-referential terms (Chisholm 1957). It seems clear, however, that representation in the subjective sense exhibits both features. Thus, the envatted brain's apple experience presents an apple to the brain even though there is no apple, and its morning-star experiences are 'presentationally different' from its evening-star experiences even though the heavenly body is one and the same.[27] Accordingly, from 'the envatted brain's experience presents an apple to the brain' we cannot truth-preservingly infer 'there is something that the envatted brain's experience presents to it' (this is failure of existential generalization), and from 'the envatted brain's experience presents the morning star to the brain' and 'the morning star is the evening star' we cannot truth-preservingly infer 'the envatted brain's experience presents the evening star to it' (substitution failure). So, subjective representation does exhibit the definitive features of intentionality, and thus qualifies as representation in the philosophically relevant sense. In other words, there *is* such a thing as *subjective intentionality*.[28]

I conclude that eliminativism is *prima facie* highly implausible. It is true that the combination of familiar tracking theories of mental representation and a compelling argument for eliminativism about subjective representation would protect these theories' status as fully adequate to the phenomena. But it is far from obvious what such an argument would look like. Arguments to the effect that subjective representation is explanatorily and descriptively dispensable, metaphysically impossible, or philosophically irrelevant do not appear to work. Another argument might, but it is the

burden of the tracking theorist to produce it. Observe that this is, essentially, the burden to supplement her tracking theory with a compelling argument for eliminativism about subjective representation.

7. Reductivism

The reductivist gambit in this area is to develop a broadly causal-covariational or teleological account of subjective representation. In doing so, the reductivist would show that familiar tracking theories of mental representation, even if not *geared* toward subjective representation, have the *resources* to account for it.

The simplest reductivist account would identify *something* that both my and my envatted duplicate's experiences track. Thus, my appe-ish experience and my duplicate's are elicited the same sensory stimulation. To be sure, unlike my brain, the envatted one is not 'coated' with a genuine sensorium, so the relevant sensory stimulation must be construed as stimulation of entry points to the brain itself – e.g., the lateral geniculate nucleus (LGN) for visual stimulation. In a way, then, both my apple-ish experiences and my duplicate's covary stably with the presence of the right kind of LGN state. A reductivist might therefore suggest that subjective apple representation is nothing but the tracking of apple-appropriate LGN states. More generally, she may attempt to account for subjective representation in terms of tracking of the right intra-cranial states.

This reductivist attempt is unsatisfactory as it stands. For presumably subjective apple representations represent apples, or apple-y things, not LGN states. It is perfectly legitimate to account for subjective representation of *x* (partly) in terms of tracking of *y*, but the reductivist does owe us an account of the relation R that holds between *x* and *y* and that enables a state to subjectively-represents *x* in virtue of tracking *y*. Clearly, R is not some other tracking relation, since the envatted brain's relevant LGN state bears no tracking relation to apples. Some other account of R would have to be provided. This further account would effectively constitute supplementation of familiar tracking-theories of mental representation.[29]

Another reductivist approach might suggest taking the best among familiar theories of objective representation and adding to it a condition of subjective

accessibility. Thus, one might hold that a mental state M presents a feature F to its subject iff (i) M tracks F and (ii) M's tracking of F is somehow introspectively accessible to the subject. If this is right, then familiar theories of mental representation only need to be supplemented with an account of introspective access to accommodate subjective representation.

The trouble with this reductive account is that it is falsified by the brain-in-vat scenario: the envatted brain has a mental state that presents an apple to it, even though it is not the case that (i) it tracks an apple and (ii) its tracking of the apple is introspectively accessible to the brain. Condition (i) is not met. If we replace condition (i) with the requirement that the state track an apple-appropriate LGN state, the problem attending the previous reductive account would reemerge: we would need an account of the (non-tracking) relation between the relevant LGN state and apples (or apple-y things).[30]

A third reductive option might attempt to account for what an experience E subjectively-represents not in terms of anything E objectively-represents but in terms of what E is objectively-represented to objectively-represent. Suppose E tracks F, but is accompanied by a higher-order mental state that objectively-represents E as tracking G. Then on this view, although E objectively-represents F, it subjectively-represents G. Naturally, the higher-order objective representation is itself accounted for in terms of tracking. The upshot is an account of subjective representation that secures its independence from objective representation but at the same time appeals exclusively to materials already used in familiar theories of mental representation (it combines familiar materials in an unfamiliar way).

This reductive approach to subjective representation is much more promising, but it does face two important challenges. First, the extant literature on naturalistic theories of mental representation includes, to my knowledge, no higher-order tracking account.[31] In this respect at least, such an account would effectively constitute *supplementation* of existing familiar theories. Secondly, and more pressingly, it is far from clear how higher-order tracking could deliver the right results in brain-in-vat scenarios. To account for the brain's experience presenting an apple to it, this reductivist would have to say that the brain harbors a higher-order state that tracks an apple-tracker. But it is not at all clear how the brain might *acquire* this higher-order tracker. To acquire such an apple-tracker-

tracker, it would have to possess a first-order apple-tracker, but (plausibly) the acquisition of apple-trackers depends on causal interaction with apples, and the envatted brain enjoys none.[32] As noted above, the only apple-relevant features with which the envatted brain's apple-ish experiences causally interact are ('apple-appropriate') LGN states. But then any higher-order state that would track these would *not* be tracking apple-trackers; it would be tracking LGN-trackers. Here again, then, the proposed reductive account appears to require supplementation, indeed the very same supplementation discussed above.

There may be other reductive accounts which have not occurred to me, and which may succeed in casting subjective representation as an ontological free lunch (given the existence of objective representation). But as with eliminativism, it is the reductivist's burden to provide any such account. In any case, a reductivist defense of familiar tracking-based theories of mental representation would require two parts: (i) an account of objective representation in terms of tracking relations and (ii) a reductive account of subjective representation in terms of objective representation. The extant literature on mental representation contains admirable work toward (i), but virtually no work toward (ii). Supplying (ii) is the minimal supplementation familiar theories would require.

8. Nonreductivism and More Radical Options

A subjective-representation enthusiast may turn the tables on the proponent of familiar tracking theories by arguing for eliminativism or reductivism about *objective* representation. Recall that one of the eliminativist tacks above involved denying that the property shared by my envatted subjective duplicate qualifies as representational in the 'philosophically interesting' sense of *intentionality*. A parallel claim could be made by the subjective-representation enthusiast. She can claim that so-called objective representation does not exhibit the features underlying failure of substitution and existential generalization.[33]

The notion that objective representation does not exhibit the feature underlying substitution failure has already been foreshadowed in the literature. Some arguments due to Searle (1991, 1992 Ch.7) and Loar (1995) could certainly be adapted in this direction. The thought is that any mental state that tracked Phosphorus would *eo ipso* be tracking

Hesperus. Tracking relations, even teleologically augmented, simply cannot discriminate between coextensive entities – for reasons explored already a generation ago (see esp. Fodor 1984). In consequence, 'M tracks F' and 'F=G' *do* entail 'M tracks G' – contrary to substitution failure.[34]

There are also arguments to the effect that objective representation does not exhibit the feature underlying failure of existential generalization (Kriegel 2011 Ch.3). Here the idea is that tracking relations, even when teleologically augmented, cannot obtain in the absence of their relata. In some worlds inhabited by my envatted duplicate there are apples, but in some there are not. Plausibly, tracking relations require the existence of the tracked. If so, in an apple-less world, my envatted duplicate's internal states could not track, hence could not objectively-represent, apples. In consequence, 'M tracks F' *does* entail 'There is an *x*, such that M tracks *x*' – contrary to failure of existential generalization.

Such arguments (which admittedly require more sustained development) might inspire some to go eliminativist with respect to objective representation. The claim would be that objective tracking relations do not qualify as representational in the philosophically interesting sense – they are not intentional.

Others may take the same considerations to suggest a *reductivist* account of objective representation. Here the idea would be that tracking relations qualify as representational (in the philosophical sense) only in virtue of bearing the right relation to subjective representation. Again following Searle's (Ibid.) lead, one might suggest that M objectively-represents F iff (i) M tracks F and (ii) M *potentially* subjectively-represents F (or M is *disposed* to subjectively-represent F). More circuitously, and now following Loar's (2003) lead, one might hold that M objectively-represents F iff (i) M tracks F and (ii) M is functionally/inferentially integrated with certain subjective representations. Other reductive accounts could also be suggested.[35] What they would all have in common is the claim that tracking, however sophisticated, only qualifies as representational in the relevant (read: intentional) sense if it bear the right relation to subjective representation. Objective representations thus inherit their status as representations *from* subjective representations. In this way, objective representation is grounded in subjective representation – without the latter there could not be the former.

The most antecedently neutral approach would be nonreductivist, rejecting eliminativism and reductivism about either objective or subjective representation. On this view, the conceptual distinction between two notions of representation is paralleled by a real distinction between two mutually irreducible kinds of representational property. Importantly, such two-way nonreductivism does constitute a departure from familiar tracking theories of mental representation, as it invites supplementation of such theories by some account of subjective representation.[36]

Such an account would tell us what it is in virtue of which a mental state M subjectively-represents a feature F. Just as certain tracking relations are designated by familiar theories as underlying objective representation, so some other relations (or perhaps monadic properties) would have to be designated as underlying subjective representation. One way to understand the recent flurry of work surrounding the so-called phenomenal intentionality research program (Kriegel 2013) is in this context: mental states represent subjectively in virtue of their phenomenal character. This is not the place to discuss work within this research program.[37] But note that insofar as the notion of subjective representation is motivated by consideration of environmentally insulated phenomenal duplicates such as brains in vats, it is *prima facie* plausible that phenomenal character would be crucial to subjective representation. Within a nonreductive framework, this is not taken to be due to the fact that tracking relations underlie phenomenal character (as in Dretske 1995); rather, tracking and phenomenology are taken to be two different sources of two different kinds of representation.

Conclusion and Future Work

It has not been my goal, in this paper, to make a case for this sort of nonreductivism, casting objective and subjective representation as mutually irreducible.[38] My goal has been more modest, and may be described as follows. Once we draw a *conceptual* distinction between objective and subjective representation, and rule out eliminativism about the latter, a certain structure emerges for a *general* theory of mental representation – a theory that accounts for *all* phenomena of mental representation. Such a general theory would comprise three chapters: (a) a theory of objective representation, (b) a

theory of subjective representation, and (c) an account of the relationship between them. My goal in this paper has been to argue that familiar theories of mental representation in terms of tracking relations are inadequate as *general* theories. For they offer us only (a), and are silent on (b) and (c). They thus fail to account for all the phenomena of mental representation. To do so, they would have to add either (i) a reductive account of subjective representation in terms of objective representation or (ii) an independent account of subjective representation. To repeat, I am open to the possibility that a reductive account of subjective representation in terms of objective representation might turn out to be right. Such an account would effectively constitute an approach to (c), and would pave the way to (b). Still, none of the familiar tracking theories of mental representation in the extant literature actually offers such a reductive account of subjective representation. It remains an outstanding intellectual debt on the part of tracking theories. My contention is that the phenomena of subjective representation cannot be dealt with simply through disregard; some positive account of them – if only a reductive account – is called for.[39]

References

- Brogaard, B. 2013. 'Do We Perceive Natural Kind Properties?' *Philosophical Studies* 162: 35-42.
- Brentano, F. 1874. *Psychology from an Empirical Standpoint*. Trans. A.C. Rancurello, D.B. Terrell, and L.L. McAlister. London: Routledge and Kegan Paul, 1973.
- Chalmers, D.J. 1996. *The Conscious Mind*. New York: Oxford UP.
- Chisholm, R. 1957. *Perceiving: A Philosophical Study*. Ithaca: Cornell UP.
- Dennett, D.C. 1969. *Content and Consciousness*. London: Routledge.
- Dretske, F.I. 1971. 'Conclusive Reasons.' *Australasian Journal of Philosophy* 49: 1-22.
- Dretske, F.I. 1981. *Knowledge and the Flow of Information*. Oxford: Clarendon.
- Dretske, F.I. 1988. *Explaining Behavior*. Cambridge MA: MIT Press.
- Dretske, F.I. 1995. *Naturalizing the Mind*. Cambridge MA: MIT Press.
- Fodor, J.A. 1984. 'Semantics, Wisconsin Style.' *Synthese* 59: 231-250.
- Fodor, J.A. 1990. *A Theory of Content and Other Essays*. Cambridge MA: MIT Press.
- Horgan, T. and J. Tienson 2002. 'The Intentionality of Phenomenology and the Phenomenology of Intentionality.' In D. J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford and New York: Oxford University Press.

- Horwich, P. 1998. *Truth*. Oxford: Oxford University Press.
- Johnston, M. 1997. 'Manifest Kinds.' *Journal of Philosophy* 94: 564-583.
- Kriegel, U. 2003. 'Is Intentionality Dependent upon Consciousness?' *Philosophical Studies* 116: 271-307.
- Kriegel, U. 2010. 'Intentionality and Normativity.' *Philosophical Issues* 20: 185-208.
- Kriegel, U. 2011. *The Sources of Intentionality*. Oxford and New York: Oxford University Press.
- Kriegel, U. 2013. 'The Phenomenal Intentionality Research Program.' In U. Kriegel (ed.), *Phenomenal Intentionality: New Essays*. Oxford and New York: Oxford University Press.
- Loar, B. 1987. 'Subjective Intentionality.' *Philosophical Topics* 15: 89-124.
- Loar, B. 1995. 'Reference from the First-Person Perspective.' *Philosophical Issues* 6: 53-72.
- Loar, B. 2003. 'Phenomenal Intentionality as the Basis for Mental Content.' In M. Hahn and B. Ramberg (eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. Cambridge MA: MIT Press.
- McGinn, C. 1988. 'Consciousness and Content.' In *Proceedings of the British Academy* 76: 219-239.
- McGinn, C. 1989. *Mental Content*. Oxford: Blackwell.
- Millikan, R.G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge MA: MIT Press.
- Millikan, R.G. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge MA: MIT Press.
- Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.
- Putnam, H. 1975. 'The Meaning of "Meaning".' In his *Mind, Language, and Reality*. Cambridge: Cambridge UP.
- Putnam, H. 1981. *Reason, Truth, and History*. Cambridge: Cambridge UP.
- Searle, J.R. 1991. 'Consciousness, Unconsciousness, and Intentionality.' *Philosophical Issues* 1: 45-66.
- Searle, J.R. 1992. *The Rediscovery of Mind*. Cambridge MA: MIT Press.
- Siewert, C.P. 1998. *The Significance of Consciousness*. Princeton NJ: Princeton University Press.
- Stampe D. 1977. 'Towards a Causal Theory of Linguistic Representation.' *Midwest Studies in Philosophy* 2: 42-63.
- Strawson, G. 2008. 'Real Intentionality 3: Why Intentionality Entails Consciousness.' In his *Real Materialism and Other Essays*. Oxford: OUP.

[1] We could, of course, use 'representation-as-tracking' and 'representation-as-presentation.' But while these terms applied intuitively in the brain-in-vat scenario, words undergo subtle changes in connotative profile

across contexts, and so these terms may not apply intuitively in other scenarios, even though there are no important substantial differences. (This is especially true of the 'presentation.') In addition, the expressions 'representation-as-tracking' and 'representation-as-presentation' are clumsy.

[2] What makes similarity relevant or sufficient is of course a complicated matter. If we had a full account of this, we could offer something like necessary and sufficient conditions for objective and subjective representation. Objective representation: '*x* is an objective representation iff *x* is similar to the envatted brain's state of tracking S1 in respect R and to degree D.' Subjective representation: '*x* is a subjective representation iff *x* is similar to the envatted brain's state of presenting an apple to the brain in respect R* and to degree D*.' For an initial grasp of the two notions, however, we can rely on intuitive takes on relevance and sufficiency.

[3] Dennett (1969) distinguishes between personal and sub-personal mental states: the former are states of us, the latter are not. This distinction applies to representational states: sometimes we do the representing, sometimes sub-systems within us do it. The distinction between wide and narrow representations is due to Putnam (1975): the former are not shared by intrinsic duplicates, the latter are. And Chalmers (1996 Ch.1) distinguishes between psychological and phenomenal conceptions of mental phenomena: the former characterizes mental phenomena in terms of their causal relations to each other and to the environment (their 'long-armed functional role'), the latter characterizes them in terms of their experiential feels (their 'phenomenal character'). This *conceptual* distinction, too, applies to mental representations as well (see Kriegel 2010).

[4] Note that at this stage no claim is made about metaphysical possibility, only about conceptual possibility. The reason is that at this stage I am only interested in the two concepts of representation.

[5] For the notion of a manifest kind, see Johnston 1997.

[6] This view is defended by Brogaard (2013), among others.

[7] Here too, it may well be possible to justifiably believe, and know, that you are not the space soul, but again this would be non-demonstrative knowledge, and being non-demonstrative, it would not *rule out* possible alternatives.

[8] It is worth stressing that it matters not whether the situation the thought-experiment enjoins us to envisage is genuinely possible or merely conceivable. For our distinction is in the first instance between two *notions*, not two *properties*. I will address the issue of whether there is a different property corresponding to each notion later on.

[9] It was pointed out to me, in this connection, that the Oxford English Dictionary's entry on representation is parceled out into seven different meanings, each sub-divided into distinct usages, in a way that suggests a heterogeneous domain of phenomena. Although the entry's authors' taxonomy and organization of the domain leaves much to be desired, the multitude and variation of meaning seems to be a genuine phenomenon.

[10] Likewise, it is not as though literary and mathematical representations represent due to the same representation relation, with the former qualifying as literary simply because they employ literary vehicles of representation and the latter as mathematical because they employ mathematical vehicles. We can imagine a single item functioning as a diagram in one context and a metaphor in another.

[11] It is important to note, however, that there is also an intimate relation between the two notions, inasmuch as the following non-trivial connection holds: whenever I have a conscious experience which represents veridically in both the objective sense and the subjective sense, what it tracks and what it presents to me are one and the same. This congruence can hardly be an accident. There is thus a non-accidental tie between what an experience represents in the objective and subjective senses when everything goes the way it

should. So while we need separate theories for mental representation in each sense, there needs to be sufficient contact or overlap between them to explain this non-accidental tie.

[12] This last condition requires that the non-Fs cause Ms because Fs cause Ms whereas the Fs cause Ms not because non-Fs cause Ms.

[13] Arguably, conditions (i) and (iii) are not satisfied either, but we do not have to worry about that, as the dissatisfaction of (ii) is sufficient to generate the problem.

[14] Moreover, in most cases (though not quite all), the nomic dependence conditions means that Fs are the only lawful causes of Ms. It is clear that this more specific condition will not be satisfied by the envatted brain's apple experiences, since they have no apples as causes, lawful or not.

[15] The term 'correspondence' is Millikan's (1984, 1993), and is left unexplained. One would be warranted to suspect that the relevant correspondence relation is ultimately to be accounted for in terms of the sort of tracking relation appealed to by causal-convariational theories. If so, teleosemantics is in fact just a teleological augmentation of causal-convariational theories.

[16] This line of argumentation against teleosemantics parallels closely Strawson's (2008), and is in fact adapted from it. Strawson argues that teleosemantics cannot accommodate the possibility of *pure observers*, because such observers are incapable of adaptive behavior. As example of pure observers, he cites the 'weather watchers': intelligent creatures stuck to the ground and unable to move but sensitive to and intensely interested in the weather. These creatures do not behave at all, let alone adaptively, but clearly have mental representations, argues Strawson. We can stipulate that something very similar is true of the envatted brain: it does not genuinely interact with its environment, though it seems to itself to do so.

[17] Recall that in our version of the tale the machine stimulates the brain randomly, it is not controlled by an intelligent and purposeful 'evil scientist.'

[18] The term 'interpreter' is used somewhat technically here, to designate any agent engaged in ascription of representational states.

[19] We can think of it this way: if objective and subjective representation were one and the same property, given that S1 is not an apple one of the two ascriptions would have to be false. Conversely, since both ascriptions are true, objective and subjective representation must be two different properties.

[20] What I mean by 'true in the minimalist sense' is something like: true in the sense of satisfying the platitudes used to formulate the so-called truth schema in minimalist or deflationary theories of truth, such as Horwich's (1998).

[21] We may also define a looser version of reductivism, allowing for reduction of subjective representation to objective representation plus other familiar and recognizably 'kosher' materials, such as functional role or other causal/mechanistic notions. I will ignore this possibility in what follows, because everything I say about the more rigorous type of reductivism should apply *mutatis mutandis* to this looser variety as well.

[22] Clearly, the proponent of familiar tracking theories of mental representation must argue for the eliminative or reductive view. If either turns out plausible, the familiar theories may be in good shape. The nonreductive view is the least conservative of the three options, and would straightforwardly require that familiar theories be supplemented with a distinct account of subjective representation. However, I will argue that even reductive views, at least in their plausible versions, would require meaningful supplementation of familiar theories of mental representation, in a sense to be duly explained. (Importantly, there are also more radical epistemic possibilities in the area. One is eliminative in the opposite direction, denying the existence of a property of objective representation. Another is reductive in that opposite direction, claiming that objective representation ultimately reduces to, or is somehow grounded in,

subjective representation. If either of these more radical options prevails, familiar tracking theories of mental representation would be cast as not just inadequate but wrongheaded.)

[23] In particular, since the overt behavior of subjects can be fully explained by citing internal states' tracking of the environment, there is no need to cite any other properties of such internal states.

[24] Thus, there are ways of construing action so that the envatted brain acts in a way that calls for positing an experience that represents an apple. Suppose we can read off the monitor that controls the vat that our envatted brain initiates motion of 'its' apparent arm in the direction of the apparent apple, performs an apparent grasping motion, and brings the apple of 'its' apparent mouth. This pattern of information we see on the monitor invites an explanation that includes the claim that the brain has a representation as of an apple.

[25] In addition, even if conceivability does not under any circumstances entail metaphysical possibility, surely it provides defeasible evidence of metaphysical possibility. If so, in the absence of defeaters we would be epistemically obliged to adopt the hypothesis that the brain-in-vat scenario is metaphysically possible.

[26] Interestingly, if we require historical continuity with Brentano's notion of intentionality, the subjective notion of representation is surely the more relevant one, as Brentano (1874 Book II Chapters 1 and 7) conceives of intentionality in terms of what presentation (*vorstellungen*) present to the subject, not in terms of any tracking relations to the environment. In fact, for most sensible properties Brentano appears to take a Kantian approach, taking them to be in some sense projected by the mind rather than inherent in the objective order of things.

[27] In saying that the experiences are 'presentationally different,' I mean to suggest that – at least for an envatted brain unaware of the identity of Phosphorus and Hesperus – what is presented in the case of one experience feels different from what is presented in the case of the other. I say more about this in Kriegel 2011 Ch.3.

[28] Loar (1987) uses the term 'subjective intentionality' to pick out something which ends up being more or less the same as *our* subjective intentionality. But my own way of introducing the notion is much more theoretically neutral: as a label for one sense of representation we find in contemplating brain-in-vat scenarios.

[29] Furthermore, nothing about this reductivist account captures the subjective dimension of subjective apple representations, the fact that they present apples *to the subject*.

[30] Furthermore, for either version of this reductivist gambit to succeed, an account of introspective access that did not appeal to subjective representation would have to be devised; this may not be straightforward. In particular, it is not obviously easier to describe introspective access without citing subjective representation than it is to describe the brain-in-vat scenario without doing so.

[31] I develop an account of this sort in Kriegel 2011 Ch.2, but naturally I disregard it here.

[32] Or at least, familiar stories about tracker acquisition appear to require this – an alternative story would have to be devised if the idea is to get around such a causal-interactive requirement.

[33] A claim in the general vicinity is made by Strawson (2008).

[34] Indeed, one might reasonably suggest that even the combination of 'M tracks F' and 'F coextends with G' entails 'M tracks G.'

[35] See Kriegel 2011 Ch.4 for relevant discussion.

[36] It is worth noting that avoiding reduction of subjective representation to objective representation, while it institutes a kind of dualism about representation, does not constitute the sort of dualism that challenges physicalism. It is perfectly consistent with subjective representation being irreducible to objective representation that it is reducible to, say, 32 Hz oscillations in the hypothalamus.

[37] For relevant work, see Loar (1987, 2003), McGinn (1988), Searle (1991, 1992), Siewert (1998), Horgan and Tienson (2002), Strawson (2008), and Kriegel (2003, 2007, 2011) *inter alia*.

[38] One problem with this view is that it is does not deliver naturalization of mental representation and may even be in tension with such naturalization (see Kriegel 2003, 2011 Ch.3).

[39] For useful comments on a previous draft, I would like to thank Stephen Biggs, Davide Bordini, David Chalmers, Allan Hazlett, and Farid Masrour. I have also benefited from presenting an earlier version of the paper at the University of Wisconsin and at conferences in Bled and Geneva. I am grateful to audiences there, in particular Gregory Bochner, Juan Comesaña, Manuel García-Carpintero, Jens Kipper, Matthew Kopec, Jack Lyons, Neil Mehta, Graham Peebles, Larry Shapiro, Allan Sidelle, and Peter Vranas.