Reliable Misrepresentation and Tracking Theories of Mental Representation

Angela Mendelovici
amendel5@uwo.ca

May 7, 2012

Abstract

It is a live possibility that certain of our experiences reliably misrepresent the world around us. I argue that tracking theories of mental representation (e.g. those of Dretske, Fodor, and Millikan) have difficulty allowing for this possibility, and that this is a major consideration against them.

1 Introduction

It is a live possibility that there are no colors. Objects appear to be colored, but upon closer examination, it could turn out that they do not have the properties our color-experiences represent them as having. If this is the case, then our color-experiences are mistaken; they *misrepresent*. Further, they misrepresent in the same way all the time. If our color-experiences misrepresent an object as red on one occasion, they are likely to misrepresent it as red on other occasions; they *reliably* misrepresent. Whether or not this is the right view of colors, it seems there could turn out to be such cases of reliable misrepresentation.

In this paper, I argue that a certain prominent class of theories of mental representation, tracking theories, make it practically impossible for mental states to reliably misrepresent, and that this is a serious problem for them.

According to tracking theories, mental representation is a relation of causation or correlation holding between mental representations and things in the world in content-endowing circumstances, e.g. circumstances in which the tokening of a representation is useful, adaptive, or involves a sufficiently strong causal connection. At least in their contemporary guise, tracking theories of mental representation emerged and gained prominence in the 1980s and 1990s, with notable developments from Fred Dretske (1981, 1988, 1995), Ruth Millikan (1984, 1989), and Jerry Fodor (1987, 1990, 1994). Despite recent interest in alternative theories of mental representation, such as disjunctivism as a theory of perception, and the phenomenal intentionality theory (Horgan and Tienson, 2002), tracking theories remain popular today.

It is well-known that tracking theories face challenges in accounting for misrepresentation, but discussion usually focuses on the **disjunction problem**: If a representation represents whatever causes its tokens (or whatever its tokens correlate with), then it seems it can never misrepresent, since anything that causes (or correlates with) any of its tokens is automatically part of its content. As will become clear, the problem of reliable misrepresentation is not the disjunction problem. Solving the disjunction problem requires allowing for *occasional* misrepresentation, while a solution to the problem I describe requires allowing for *systematic* misrepresentation, and it turns out that, given the tracking theory's resources, allowing for occasional misrepresentation is much easier than allowing for systematic misrepresentation.

This paper proceeds as follows: §§2–3 describe reliable misrepresentation and tracking theories of mental representation. §4 argues that tracking theories face difficulties in allowing for reliable misrepresentation. §5 offers a diagnosis of this difficulty. §6 argues that the difficulty in allowing for reliable misrepresentation is a serious problem for tracking theories.

2 Reliable misrepresentation

In statistics, there is a distinction between a test's validity and a test's reliability. A **valid** test is one that fairly accurately detects what it is intended to detect. A **reliable** test is one that yields more or less the same results each time it is administered, regardless of whether it is valid. We need this distinction because it turns out that we can have reliable tests that are not valid. For example, it is sometimes claimed that the Standard Aptitude Test (SAT) is quite bad at predicting success in college, which is what it is supposed to predict, and so that it is an invalid test. However, it is generally agreed that the test is reliable in that it yields more or less the same results when administered to the same subjects on distinct occasions.

I claim that we need a similar distinction when it comes to mental representation, in this case between **reliability** and **veridicality**. We need this distinction because there might be representations that **reliably misrepresent**. They are *reliable* in that they respond similarly in similar circumstances, but they *misrepresent*, since the world isn't really as they represent it to be. Loosely, reliable misrepresentation is getting things wrong in the same way all the time.

Before offering a more precise characterization of reliable misrepresentation, some definitions are in order. **Mental representation** is the *aboutness* of mental states. What a mental state or mental representation is about is its **content**. The types of representational states I am interested in are those representing properties, roughly understood as *ways things are or might be*. The metaphysical status of properties is left open: they might be universals, tropes, or something else. Mental states involving the representation of properties might have singular contents, such as that object o has property P, or they might have existential contents, such as that there is an object x that has property P. For

simplicity, I will sometimes write as if mental states have singular contents, but everything I say holds if we take the mental states in question to have existential contents instead.

The kind of mental states that most uncontroversially are candidates for misrepresentation are states that in some sense "say" that some represented object has some represented property. Call such states **attributive** states. Possible examples include perceptual experiences and beliefs. Unlike desires and hopes, these states are assessable for accuracy, and thus can misrepresent. The claim that a representation R reliably misrepresents, then, should not be understood as the claim that for the most part, tokens of R are involved in states that misrepresent, since many of R's tokens might occur in mental states that are not assessable for accuracy, such as desires. Rather, the claim that R reliably misrepresents should be understood as the claim that *attributive* mental states involving tokens of R are usually false or nonveridical.

We are now in a position to offer a more precise characterization of reliable misrepresentation. An organism's representation of type R **reliably misrepresents** some property P if and only if

(1) Some tokens of R are involved in attributive mental states that represent objects as having property P,

(2) Most or all of the relevant objects do not have P,

(3) Tokens of R do or would nonveridically represent objects as having P in the same types of circumstances on separate occasions.

(1) and (2) yield misrepresentation, while (3) yields reliability.[1] For example, suppose the scenario described in the beginning of this paper is actual, and

[1] The definition is a little vague, since we haven't said anything about when circumstances count as being of the same type. But we can allow reliability to also be a vague or graded notion. None of this affects my argument.

objects do not have the color properties our visual experiences represent them as having.[2] Since conditions (1) and (2) are satisfied, and color-representations misrepresent. Further, they misrepresent in the same way in similar circumstances. If you misrepresent a tomato as red on one occasion, you are very likely to misrepresent it as red on future occasions. Color-representations thus satisfy (3), and so they *reliably* misrepresent.

Reliable misrepresentation usually involves tracking. For now, we can take tracking to be characterized by the intuitive notion of keeping track of, indicating, or carrying information about something; the next section characterizes tracking more precisely. In cases of reliable misrepresentation, there is usually some property that is causally related to the representation in question, and that is thus successfully tracked. Since reliable misrepresentation is misrepresentation of the same sort in certain circumstances C, C is automatically tracked. More interestingly, since reliable misrepresentation is reliable, it is likely that some specific feature present in C is causally responsible for the repeated tokening of the representation in question in just those circumstances, and is thus tracked. Compare this to the SAT, which, let us assume, is reliable but invalid. The SAT tracks *something*—that's why test-takers' scores are more or less constant across multiple applications of the test—but what it tracks is not scholastic aptitude. Perhaps it tracks hours of preparation, parents' income, or other features of test takers. Likewise, a reliable misrepresentation tracks *something*, but what it tracks is not what it represents.

At this point we can further characterize validity and reliability: Valid tests are tests that track what they are supposed to test, while reliable but invalid tests are tests that track something other than what they are supposed to test. Similarly, generally veridical mental representations are mental representations that represent the same thing that they track, while reliable misrepresentations

[2] See, e.g. Maund (1995), Pautz (2006), Chalmers (2006) and Mendelovici (2010).

are representations that represent one thing and track something else. Of course, this does not automatically exclude the possibility that reliable misrepresentation involves multiple tracking relations, at least one of which holds between the representation in question and what it represents.

As should already be clear, reliable misrepresentation is not the same thing as **hallucination**. One feature of hallucination is that it is not stimulus-bound, that is, that it does not occur reliably in response to external stimuli. Hallucinations are fringe cases analogous to occasional glitches in the SAT grading process due to, say, a one-off computer malfunction. Thus, unlike reliable misrepresentation, hallucination is not reliable.[3] Additionally, hallucinations are compatible with the overall veridicality of the representation in question, just as the occasional SAT grading glitch is compatible with the test being for the most part valid, but reliable misrepresentation is not compatible with the overall veridicality of the representation in question.

Reliable misrepresentation should also be distinguished from **illusion**. Like reliable misrepresentations, illusions arise regularly and predictably, and hence are reliable. But, like hallucination, illusions are compatible with the overall veridicality of most attributive uses of the representations in question. Illusions can be understood as unintended side-effects of otherwise veridical and perhaps even optimally designed systems. They are analogous to the systematic distortion of the SAT results of test takers for whom English is a second language. This kind of distortion is possible even for a generally valid test, one that gets most cases right most of the time, and even for an optimal test, one that is as good as possible given its constraints (for instance, a constraint on the SAT is that it must be delivered in a language and not be too costly to administer). Likewise, illusions are compatible with the overall veridicality of

[3]Could there be reliable hallucinations? If so, they would count as reliable misrepresentations.

a well-functioning system. While reliable misrepresentation can be part of a well-functioning and even optimally designed system, it is not compatible with the overall veridicality of attributive uses of the relevant representations.

Finally, reliable misrepresentation should be distinguished from **occasional misrepresentation**. Occasional misrepresentation is the occasional mistaken application of a representation to an object. To borrow one of Jerry Fodor's (1987) examples, one might mistakenly identify an overweight horse in the distance as a cow. Like hallucinations, occasional misrepresentation differs from reliable misrepresentation in that it is a type of one-off occurrence that is compatible with the overall veridicality of attributive uses of the representation in question.

In this paper, I will not argue that there are actual cases of reliable misrepresentation.[4] Instead, I will argue that tracking theories face difficulties allowing for the possibility of reliable misrepresentation, and that this is a problem for those views because we should not rule out *on the basis of our theory of mental representation* that there are such cases.

3 Tracking theories of mental representation

According to **tracking theories of mental representation**, mental representation is a matter of detecting, carrying information about, or otherwise correlating with states of the environment. On some tracking views, the relevant tracking relation between a mental representation and what it represents is a **causal relation**. For example, on such views, the mental representation TIGER gets to represent *tiger* because tigers cause the tokening of TIGER in the appropriate circumstances. Of course, this causal relation between tigers and TIGER might be mediated by other causal relations, for example, a causal rela-

[4] I have argued for this in Mendelovici (2010, Ch. 5).

tion between tiger stripes and certain states of the retina, as long as TIGER and tigers are themselves connected by some causal chain.[5]

There are also tracking views on which the tracking relation is not a causal relation. If F is correlated with G (perhaps because some third item H causes both F and G, or perhaps even due to a pre-established harmony between F and G), then another way to track F is by being causally sensitive to G. Depending on what F and G are, it might be cheaper or easier to develop a causal sensitivity to one rather than the other. For example, migratory birds non-causally track certain geographic locations by causally tracking magnetic fields correlated with those locations.

As the example of migratory birds shows, there are many tracking relations obtaining between mental representations and items in the world. For instance, a migratory bird might have a representation that bears one tracking relation to Florida, and another tracking relation to geomagnetic south. But our representations do not represent everything they can be said to track. Much of the debate among tracking theorists, then, is over which tracking relation is The Representation Relation.

One reason for specifying exactly which tracking relation is to be identified with mental representation is to allow for occasional misrepresentation. If a representation represents everything it can be said to track, then it cannot misrepresent, since whatever causes or correlates with it automatically counts as part of its content on some tracking relation or other, and so the allegedly misrepresenting state will turn out to be veridical.[6] The general strategy for al-

[5] Proponents of causal tracking theories include Fred Dretske (1995) and Michael Tye (2000).

[6] In other words, a tracking theory of mental representation must avoid the disjunction problem (Fodor, 1987, Ch. 4), where a theory suffers from the disjunction problem when it wrongly counts cases of misrepresentation as cases of veridical representation of disjunctive contents. For example, if a tracking theory claims that representations represent whatever causes them in conditions C, and in conditions C a wallaby causes a kangaroo-representation, the theory wrongly counts this as a case of veridical representation of a wallaby $\vee$ kangaroo.

lowing for occasional misrepresentation is to distinguish the possible and actual tokenings of a representation that determine its content from those that do not. Tokenings of a representation that determine its content are those that occur in what we might call the **content-endowing** conditions. The tracking relation to be identified with mental representation, then, is the relation that obtains in content-endowing conditions.

There are various options for how to specify a representation's content-endowing conditions. **Optimal-functioning** or well-functioning tracking theories of mental representation take the content-endowing conditions to be conditions of optimal functioning or well-functioning, that is, the conditions in which the mental state in question now helps (or would help) its current bearer survive or flourish.[7]

Teleological tracking theories take content-endowing conditions to be design conditions, where **design conditions** are the conditions in which the tokening of a representation helped our ancestors survive and reproduce. So if food triggered R in our ancestors and this helped them survive and reproduce, then R represents *food*.[8]

Another approach is the **asymmetric dependence** theory (Fodor, 1987): A representation represents whatever causes its tokens (in a law-like way) such that for anything else that causes its tokens, the latter causal connection is dependent on the former and the former causal connection is not dependent on the latter. Dependence is cashed out counterfactually: R represents *food*

[7]Tye (2000) holds something like an optimal-functioning theory, though he also invokes teleological elements.

[8]See e.g. Millikan (1989) and Dretske (1995)). Dretske develops the view that representations represent whatever it is their function to indicate, where for some representations, this function is determined by evolution, while for other representations, it is acquired through experience. On Millikan's view, representations represent whatever conditions obtained in an organism's ancestors' environment that allowed for the use of representations to be helpful to survival. In the case of food, these conditions might include food's being a good source of nutrients. So, on this view, R represents *good source of nutrients*, which has no causal effect on the representation FOOD.

just in case food causes tokens of R and for anything else, T, that causes tokens of R, T would not cause tokens of R unless food did, whereas food would cause tokens of R even if T did not. Rather than directly distinguishing between content-endowing and non-content-endowing *conditions*, the asymmetric dependence view distinguishes between content-endowing and non-content-endowing *relations* between representations and what causes their tokens. Since I will be discussing the asymmetric dependence view alongside the above views, it will be convenient to use similar language in describing it. Thus, for the asymmetric dependence theory, let content-endowing conditions be the conditions of a representation's being tokened as a result of a law-like causal relation obtaining between a representation R and a property P such that all other causal relations between R and other properties are asymmetrically dependent on the causal relation between R and P. Intuitively, the content-endowing conditions are those in which a representation is tokened as a result of a comparatively strong causal connection.[9]

Before continuing, there is an important caveat: Tracking theories need not claim that all representations get their content from tracking. For instance, a possible view is that all simple representations get their contents from tracking, while composite representations get their contents compositionally from simple representations. For the remainder of this paper, I will be concerned only with representations that are supposed to get their content from tracking.

[9]In Fodor (1990), Fodor proposes an additional requirement that must be met in order for a representation R to represent some content P: Some instances of R must have actually been caused by P. Since in cases of reliable misrepresentation, quite plausibly *no* instances of the representation in question was caused by instances of the property that it represents, this version of the asymmetric dependence view is even easier to argue against. For this reason, my discussion focuses on a version that does not endorse this extra commitment.

4 The problem for tracking theories

For every tracking theory, there are conditions in which it is impossible to misrepresent, either because they are content-endowing conditions, or because they are of the same type as content-endowing conditions. This makes it difficult for them to allow for reliable misrepresentation, since, as I will argue, the most natural cases of reliable misrepresentation tend to involve misrepresentation in such conditions. The only way for a tracking theory to allow for reliable misrepresentation is for it to maintain that the relevant conditions do not obtain for the representation in question, which is implausible.

To argue for the general point that reliable misrepresentation is problematic for tracking theories, let us consider the situation for the types of tracking theory discussed above. I will outline the argument for each case in general terms. Then I will illustrate my argument using hypothetical cases of reliable misrepresentation. The hypothetical cases are merely provided as illustrations; nothing hangs on the cases being as I describe.[10]

4.1 Optimal-functioning theories

On an optimal-functioning theory, the content-endowing conditions are the conditions in which a representation's tokening helps its possessor survive or flourish. This theory does not allow for misrepresentation in conditions in which a representation's tokening helps one survive or flourish, since what a representation corresponds to in those circumstances just *sets* or *determines* its content.

The problem is that it seems not only possible, but fairly likely, for reliable misrepresentations to help their possessors survive and flourish. For certain tasks, survival and flourishing might only require reliability, not veridicality.

[10] Some of my arguments in this section are similar to those presented in Holman (2002), which argues that there is a tension between naturalism about mental representation and color eliminativism.

For instance, a task that might be important for survival and flourishing might involve the **re-identification** of objects through time. Suppose a representation R reliably misrepresents an object o as having a property P. Since R misrepresents, it is not veridical. However, since it *reliably* misrepresents, the possessor of R can still use R to re-identify o on multiple occasions. Every time the possessor represents the content P, she can infer that o is likely to be present. Thus, it is implausible that no cases of reliable misrepresentation could be helpful for survival and flourishing. But that is what the optimal-functioning version of the tracking theory requires. It is important to emphasize that it is *because* R is reliable that it can be useful despite misrepresenting.

To illustrate the argument with a concrete example, suppose color-experiences reliably misrepresent. The optimal-functioning theory must maintain that no actual uses of color-representations contribute to survival or flourishing (but, perhaps, that there are merely possible uses that would contribute to survival and flourishing). However, this is implausible. For instance, even if objects do not in fact have colors, we can use the colors we misrepresent them as having to re-identify them over time. For example, if you reliably misrepresent your car as red, you can quickly find your car in the parking lot by scanning your surrounds for redness. As long as you misrepresent your car as having the same color on most occasions, you can use your nonveridical representation to help you re-identify it on separate occasions. Thus, in the hypothetical scenario in which color-representations reliably misrepresent, it would be implausible to suppose that they do not contribute to survival and flourishing. But that is what the optimal-functioning tracking theory would require.

Reliable misrepresentations are also helpful for tasks requiring **discrimination**. Suppose objects of type T_1 are helpful for survival and flourishing, while objects of type T_2 are harmful for survival and flourishing. Reliable misrep-

resentation involves misrepresenting in the same way on various occasions. If objects of type T_1 are reliably misrepresented as having property P_1, while objects of type T_2 are reliably misrepresented as having property P_2, then we can use our representations of P_1 and P_2 to differentially guide our behaviors towards objects of types T_1 and T_2. What matters for survival here is that T_1 and T_2 are represented *differently*, and this can occur even if neither is represented veridically. Thus, it is quite plausible that cases of reliably misrepresentation aid in tasks requiring discrimination, and so that they generally aid in survival and flourishing.

To switch to a different illustrative example, suppose certain gustatory representations reliably misrepresent. We experience some objects as sweet and others as bitter, but in fact, these objects do not have the properties of *sweetness* and *bitterness*. Instead, objects represented as sweet have the property of containing a sufficiently large amount of sucrose, fructose, and other sugars and objects represented as bitter have the property of containing one of a large number of compounds, many of which are toxic. Even though objects do not really have the properties of sweetness and bitterness, we represent objects high in the various sugars differently from objects that are toxic. This difference in representation of the sugary from the toxic helps differentially guide our behaviors towards sugary and toxic items, and this helps us survive and flourish. What's relevant for survival and flourishing in this case is that sugary items are represented *differently* from toxic items, and this arises due to the *reliability* of the distinct misrepresentations, and despite their general non-veridicality. In conclusion, in this hypothetical scenario, it would be implausible to suppose that the reliable misrepresentation of sugary things as sweet and toxic things as bitter

does not contribute to our survival and flourishing, as the optimal-functioning tracking theory would require.[11]

In short, while the optimal functioning theory can allow for some cases of reliable misrepresentation, the cases they allow for are somewhat unnatural. They are cases in which reliability doesn't play its usual roles of allowing for re-identification and discrimination. Put otherwise, the optimal-functioning theory cannot allow for paradigm or clean cases of reliable misrepresentation, where **clean** cases are cases that exhibit the likely features of reliable misrepresentation. Since mere reliability tends to confer usefulness, one such likely feature is usefulness.

To summarize, for certain tasks reliability is sufficient to aid in survival and flourishing, and thus, reliable misrepresentations can be quite useful. But the optimal-functioning tracking theory can't allow for useful reliable misrepresentations. Perhaps there are or could be cases of reliable misrepresentations that are not useful, but it is unclear what those cases would look like, and that is precisely because mere reliability can be so useful. Put otherwise, the optimal-functioning theory cannot allow for clean cases of reliable misrepresentation, and this is severely limiting for the theory.

4.2 Teleological theories

On the teleological tracking theory, the content-endowing conditions are design conditions, that is, ancestral conditions in which the triggering of a representation helped our ancestors survive and reproduce. According to the teleological theory, content lags behind the determiner of content: what a representation

[11] Of course, the tracking theorist can respond to any putative case of reliable misrepresentation by denying that it is a case of misrepresentation. This objector does not disagree with my main claim in this section, which is that the tracking theory has trouble allowing for reliable misrepresentation. Rather, the objector might disagree with my claim in §6 that it is inappropriate for a theory of mental representation to rule out the possibility of such cases of reliable misrepresentation.

represents at time t depends on what it co-occurred with at some time in the past *t-1* such that this co-occurrence was useful for its possesser's survival and reproduction, where what determines the temporal distance between *t-1* and t is up to the theory to decide.[12]

As discussed in the previous subsection, reliable misrepresentation is useful for re-identification and discrimination, which results in improved survival and flourishing. This gives rise to a problem in allowing for reliable misrepresentation on the teleological theory: Suppose R reliably misrepresents; it falsely or inaccurately represents P but tracks a distinct property Q. If our representation R occurs in the same types of conditions as our ancestors' representation R, then the arguments from the previous subsection transform fairly straightforwardly into an argument against the teleological tracking theory: R's tracking property Q is useful for survival and reproduction in us, and so it was likely to be similarly useful to our ancestors, since R occurred in similar circumstances in us and our ancestors. Further, R doesn't occur in the presence of P in us, so it doesn't occur in the presence of P in our ancestors, since, again, R occurred in similar circumstances in us and our ancestors. But then R represents Q and not P in us, since it is R's co-occurrence with Q and not with P that helped

[12]I classify Dretske as endorsing a teleological theory based on his *Naturalizing the Mind* (1995). However, in *Explaining Behavior* (1988), he develops a non-evolutionary view on which a mental representation's function is determined by ontogenetic factors. Roughly, the view is that sometimes an internal state that is causally sensitive to an environmental feature P leads to a behavior that confers a reward. When this happens, the link between the internal state and the behavior is reinforced, and the internal state comes to acquire the function of causally indicating P, and thus comes to represent P. This view fairly straightforwardly faces problems in allowing for reliable misrepresentation. In order to acquire the function of indicating P, an internal state R must at some point have been caused by an instance of P, and this must have led to reward-conferring behavior. In cases of reliable misrepresentation, P is unlikely to have caused tokens of R at all. And since mere reliability is useful, what R tracks is likely to have caused tokens of R in the cases in which R led to reward-conferring behavior. In any case, since in allowing for misrepresentation, this version of Dretske's view exploits the time lag between the content-endowing conditions and instances of misrepresentation, much of the discussion in this section can be adapted to apply to it. In particular, the view can exploit the time lag strategy described in the main text in order to allow for certain kinds of cases of reliable misrepresentation, but these resulting cases are not clean cases.

our ancestors survive and reproduce. But then, since R veridically represents Q, and since R doesn't represent P, R doesn't reliably misrepresent P.

The teleological theory can allow for reliable misrepresentation by claiming that our circumstances have relevantly changed from those of our ancestors. At *t-1*, R co-occurred with P and this used to be helpful for survival and reproduction, but now at time t, R co-occurs with Q, and this is useful for survival and reproduction. This strategy for allowing for reliable misrepresentation exploits the time lag between the events *determining* a representation type's content and a representation type's *having* of that content. Unfortunately, this strategy makes reliable misrepresentation unstable: tracking tends to help with survival and reproduction, and thus tracked properties have a tendency towards becoming represented properties. Suppose the improbable circumstances described here obtain and R reliably misrepresents P in us. R is unlikely to continue to represent P in our descendants at time *t+1*. This is because our descendants' representations of type R represent whatever our representation R co-occurred with such that this co-occurrence was useful for our survival and reproduction. But that's not P; that's Q. Thus, the teleological theory predicts that the unlikely situation in which R reliably misrepresents P is unstable; in subsequent generations, R will come to represent Q instead.

By the same token, allowing for reliable misrepresentation by claiming that circumstances have changed doesn't allow for both us at t and our ancestors at *t-1* to reliably misrepresent. In order for our representation R to reliably misrepresent, our ancestors' representation R must have co-occured with P. This means that if our ancestors' representation R also represented P, it did not *mis*represent. Thus, if our representation R reliably misrepresents, our ancestors' representation R doesn't. If instead our ancestors' R reliably misrepresented P, then we wouldn't reliably misrepresent P, because, like our descendants in

the scenario described above, our representation R's content would be whatever it usefully tracked in our ancestors, which is Q. In other words, R's content would have had a chance to catch up with what it tracked.

We can illustrate the argument by considering again the hypothetical example of the reliable misrepresentation of color-experiences. An account of such reliable misrepresentation that exploits the time lag between content determination and the having of content would claim that our ancestors' world was colored. Our ancestors had inner states that at least sometimes corresponded to colors, and this was useful for their survival and reproduction. Since our ancestors' time, the world ceased to be colored, but we still have the same type of inner states, which now misrepresent colors. Further, since the time of our ancestors, our inner states came to track surface reflectance properties, so they *reliably* misrepresent. While our inner states reliably misrepresent colors, the same states in our descendants will veridically represent surface reflectance properties.

While it is possible to have cases of reliable misrepresentation on the teleological theory, these cases are not clean cases. They require a change in environment and are unstable. But since reliable misrepresentation is *reliable*, in clean cases of reliable misrepresentation, our representations are useful for our survival and flourishing. It's likely that this general predicament would also have been useful for our ancestors, and it is this that explains why we have them today. Since our reliable misrepresentations are useful, our descendents are likely to continue to reliably misrepresent in the same way. Reliable misrepresentation, thanks to its usefulness, tends to be evolutionarily stable.

In summary, while the teleological theory allows for cases of reliable misrepresentation in which our ancestors' tracked the represented property but circumstances have changed such that we no longer track that property, it cannot

allow for clean cases of reliable misrepresentation, and this is severely limiting for the theory.

4.3 Asymmetric dependence

According to the asymmetric dependence theory, in order for R to represent P, it must be the case that for all properties Q that are distinct from P, the Q-to-R connection is asymmetrically dependent on the P-to-R connection. One way to unpack this is as follows:

(1) If Ps didn't cause (as a matter of a law-like connection) R, then Qs wouldn't cause R either.

(2) If Qs didn't cause R, then Ps would cause R.

One way to specify the truth-conditions of these counterfactuals is in terms of possible worlds:

(1′) In the nearest possible world in which Ps don't cause R, Qs don't cause R.

(2′) In the nearest possible world in which the Qs don't cause R, Ps do cause R.

Now suppose R reliably misrepresents P. In §2, we saw that in cases of reliable misrepresentation, there is at least one property that is more or less reliably tracked by the representation in question. Let Q be such a property. Now, let us evaluate (1′). The nearest possible world in which the Ps don't cause Rs is the actual world (since, by hypothesis, R reliably *misrepresents*). But Qs do cause R in the actual world. So, (1′) is false, and the asymmetric dependence relation does not obtain in the case of reliable misrepresentation.

Thus, at least this reading of the asymmetric dependence theory does not allow for reliable misrepresentation.

It will help to consider as an illustration the hypothetical case of reliable misrepresentation in color-representation: color-representations represent uninstantiated colors but are caused by surface reflectance properties. For the asymmetric dependence theory to allow color-representations to reliably misrepresent it must be the case that the surface-reflectance-property-to-color-representation connection is asymmetrically dependent on the color-to-color-representation connection. If the unpacking of (1) and (2) in terms of possible worlds is right, then that means that in the nearest possible world in which colors do not cause color-representations, surface reflectance properties do not cause color-representations either. But the nearest possible world in which colors do not cause color-representations is *our* world (since, by hypothesis, our color-representations *misrepresent*), and surface reflectance properties do cause color-representations in our world (since, by hypothesis, our color-representations misrepresent *reliably* because they track surface reflectance properties). And so, the surface-reflectance-property-to-color-representation connection is *not* asymmetrically dependent on the color-to-color-representation connection, and at least this reading of the asymmetric dependence view cannot allow for such a case of reliable misrepresentation.

One might suggest that (1′) and (2′) are not the correct unpackings of (1) and (2). After all, Fodor claims that mental representations can bear the relevant kind of robust law-like causal connections to uninstantiated properties.[13] For example, it might be true that unicorn horn causes (in a law-like way) poisoned water to be potable, even though there are no actual instances of this causal connection. Likewise, even if P is uninstantiated, there might be a law-like causal connection between P and R. Thus, the asymmetric dependence theorist might

[13]See Fodor (1987, pp. 163–164) and (1990, pp. 100–101).

account for reliable misrepresentation by claiming that there is an uninstantiated causal connection between P and R and any law-like causal connection between Q and R is asymmetrically dependent on the P-to-R connection.

To assess this possibility, it will help to consider the general idea behind (1) and (2). The general idea is that the relation between a representation and what it represents is somehow stronger than the relation between the representation and its other causal triggers. It does not matter whether one of those relationships is not instantiated in the actual world, because an uninstantiated connection can be stronger than an instantiated one. Unfortunately, however, the kind of senario Fodor would need to obtain seems unlikely in the case of reliable misrepresentation. The problem is that reliable misrepresentation is *reliable*, and what's responsible for its reliability is precisely the representation's connection to something other than what it represents. In the general case described above, what accounts for the reliability of R's misrepresentation of P is precisely its connection to Q. And so the connection between Q and R is likely to count as fairly strong on any unpacking of what is meant by "strength". Not only is the Q-to-R connection fairly strong, but it is also not clear that any P-to-R connection, say, obtaining in some other world, would likewise be strong. If R is a case of reliable misrepresentation, then it is likely to be hooked up with our other states and features of the environment such that it is a good detector of Q. It is doubtful that in an alternate possible world in which P was instantiated the very same representation R would be causally sensitive to P instead of or as well as Q.

In sum, the problem is that reliability usually implies a relatively strong causal relation, but on the asymmetric dependence theory, misrepresentation requires a relative failure of strength. Thus, on the asymmetric dependence theory, reliability is at odds with misrepresentation, making it difficult to allow

for reliable misrepresentation. Put otherwise, since reliability usually implies a relatively strong causal relation, the only kinds of cases of reliable misrepresentation that the asymmetric dependence theory can allow for are fairly unclean.

It will help to return to the hypothetical example of color. As suggested for the general case above, one might insist that in the actual world, the color-to-color-representation connection exists, but is uninstantiated, and further, that the surface-reflectance-property-to-color-representation connection is asymmetrically dependent on it. But the surface-reflectance-property-to-color-representation connection is fairly strong. It is at least as strong as the horse-to-HORSE connection, which Fodor seems to consider a paradigm example of a strong connection (see Fodor (1987, Ch. 4)). That means that in order for the asymmetric dependence theorist's strategy to work, the uninstantiated color-to-color-representation connection must be still more strong than this, which is implausible. One way to see the implausibility of this account of the reliable misrepresentation of colors is to compare a natural explanation of the causation of color-experiences to the explanation offered by the present account. The natural explanation is this: There is a law-like causal connection between surface reflectance properties and color-representations. The explanation the asymmetric dependence theory must give is this: There is an uninstantiated law-like causal connection between colors and color-representations. There is also an instantiated causal connection between surface reflectance properties and color-representations, but this connection "hijacks" or is otherwise dependent on the previously mentioned uninstantiated connection between colors and color-representations. This second candidate explanation appealing to uninstantiated laws is needlessly complex and implausible, and receives no theory-independent motivation. Causal explanations of other cases of reliable misrepresentation are likely to have the same structure. Thus, even though there is a way to accommo-

date reliable misrepresentation on the asymmetric dependence view, it is quite unappealing.

Further, it is not clear that in a world in which there were colors as we represent them, colors would cause our color-representations. Why should we think that color-representations that have evolved to be sensitive to surface reflectance properties would in such a world be causally sensitive to colors?

In summary, in cases of reliable misrepresentation, there is a strong connection between a representation and something that it does not represent, and that makes it difficult for the asymmetric dependence theory to allow for reliable misrepresentation. The only way for the asymmetric dependence theory to allow for reliable misrepresentation is to insist that there is an uninstantiated law-like causal connection on which the instantiated law-like causal connection is asymmetrically dependent, but this is quite implausible and contrived, and does not allow for clean cases of reliable misrepresentation.

4.4 Taking stock

In this section, I have argued that the prominent existing types of tracking theories have difficulty allowing for reliable misrepresentation. The kinds of reliable misrepresentation they allow for lack the usual concomitants of mere reliability that are present in clean cases, including usefulness, stability, and strength of causal connection.

Of course, there are other possible tracking theories, but it is difficult to imagine one that can avoid these difficulties. As I will argue in the next section, the feature of reliable misrepresentation that generates the problem is that apart from being non-veridical, reliable misrepresentations are otherwise very well-behaved.

5 Diagnosis of the difficulty

Why do tracking theories have difficulty allowing for reliable misrepresentation? The problem is that tracking theories peg veridicality to their favored notion of **nonsemantic success**, a type of success distinct from veridicality. A state is nonsemantically successful when it occurs in conditions such as conditions of optimal functioning, conditions of the same type as the design conditions our ancestors found themselves in, or, for the asymmetric dependence theory, when it is an instance of a relatively strong connection. The connections a mental representation has in content-endowing conditions determine its content, and nonsemantically successful conditions are conditions either identical to or of the same type as content-endowing conditions.[14] As a result, a representation cannot misrepresent in nonsemantically successful conditions. But that means that whenever there is misrepresentation, there must be a **nonsemantic defect**, a defect apart from being nonveridical.

Appeal to nonsemantic defects allows tracking theories to handle misrepresentation in cases that are nonsemantically defective. They can deal with hallucinations, occasional misidentifications of malnourished cows as horses, and illusions that occur in circumstances that a representational system did not specifically evolve to handle. Since all those cases plausibly involve nonsemantic defects, they are excluded from the types of cases that determine the content of the representations in question, and they can be correctly classified as misrepresentations by comparing their causes to the causes of the same representations in nonsemantically successful conditions.

Reliable misrepresentation is unique among types of misrepresentation in that it needn't be accompanied by a nonsematic defect. It can occur in condi-

[14]For the optimal-functioning theory and the asymmetric dependence theory, nonsemantically successful conditions just are content-endowing conditions, while for the teleological theory or any historical theory, nonsemantically successful conditions are conditions of the same type as the content-endowing conditions.

tions of optimal functioning, conditions of the same type as design conditions, and as a result of a robust causal connection. As a result, the requirement that nonsemantic defects accompany misrepresentation makes reliable misrepresentation practically impossible on tracking theories. Indeed, we've seen that the only cases that they can allow for are unclean cases, and unclean cases are cases in which reliable misrepresentation happens to be accompanied by a nonsemantic defect. Clean cases, on the other hand, do not involve nonsemantic defects, and that is because mere reliability tends to confer all the relevant nonsemantic virtues, such as usefulness, stability, and a relatively strong causal connection.[15]

6 Why we should allow for reliable misrepresentation

Of course, I have not argued that there are any cases of reliable misrepresentation. So why is it a problem that tracking theories are ill-suited to allow for them? The problem is that whether or not there are such cases, it would be inappropriate to conclude that there aren't *on the basis of a metaphysical theory of mental representation.* By a **metaphysical theory of mental representation**, I mean a theory that aims to tell us what mental representation *really is*, as opposed to a theory that tells us certain further facts about mental representation, such as facts about the structure of various representational spaces, which specific contents we represent, or whether any particular representation is veridical.

To see how strong the tracking theorist's commitment to there being no cases of reliable misrepresentation in ideal conditions really is, notice that dis-

[15]Tracking theories' connection between semantic and nonsemantic success is related to the goal of *reducing* facts about mental representation to other naturalistically-acceptable facts. In the next section, I will argue that a theory of mental representation should allow for the possibility of reliable misrepresentation. If that is right, then perhaps we should look for a reductive account of mental representation elsewhere.

junctivism, a theory that treats veridical representation and hallucination differently, does not incur these kinds of empirical commitments: According to disjunctivism, there is one story about how an experience gets to represent if it is veridical, and a different story about how an experience gets to represent (or appear to represent) if it is hallucinatory. But disjunctivism nonetheless leaves it an open empirical question whether any *particular* experience (or appearance of an experience) is veridical, hallucinatory, or misrepresents in some other way. In contrast, tracking theories close off the empirical possibility of clean cases of reliable misrepresentation.

I think it is immediately clear that we should want to leave open the possibility of clean cases of reliable misrepresentation. That we have states that reliably misrepresent in this way is a genuine empirical possibility, and one that I think is likely to be actual in at least some cases. However, for those not yet convinced, the following two subsections outline two specific reasons why we should want a metaphysical theory of mental representation to allow for reliable misrepresentation, one having to do with psychological explanation, the other having to do with our epistemic position regarding metaphysical facts.

6.1 Psychological explanation

A theory that allows for reliable misrepresentation has more explanatory resources than one that does not. In particular, such a theory makes room for two distinct explanatorily relevant features of mental states: reliability and veridicality. Distinguishing these two features allows for certain kinds of psychological explanations of patterns of behaviors and responses that would otherwise be foreclosed. In particular, it allows for an appealing explanation of patterns involving successful re-identification and discrimination but mistaken inferences.

Suppose an organism uses a representation R with content P to successfully re-identify objects over time and to distinguish objects that trigger R from those that do not trigger R. However, it reasonably but mistakenly infers that an object has property Q on the basis of its having property P. One attractive explanation of the pattern of behaviors and responses is that R reliably misrepresents certain kinds of objects as having P. The fact that R is *reliable* explains why the organism can use R for re-identification and discrimination. The fact that R *misrepresents* explains why some inferences involving R are unsuccessful: An inference from an object's having P to its having Q can fail when the object doesn't actually have P, even if the inference is otherwise reasonable (i.e. even if there is the supposed connection between P and Q).

Consider the following illustrative example. We can use representations of heaviness to re-identify particular objects over time (e.g. a dictionary in a dark room) and to discriminate between different types of materials by lifting them. However, our heaviness-representations are sometimes involved in unsuccessful inferences. For example, from a particular dictionary's being heavy, we might infer that it will be hard to lift on the moon.

Reliable misrepresentation offers an appealing explanation of this pattern of behaviors and responses: Representations of heaviness reliably misrepresent. They nonveridically represent objects as having an intrinsic property of heaviness, but they track relational properties obtaining between objects and other objects, e.g. the Earth or the Earth's gravitational field. The pattern of behaviors and response to be explained has two parts: (1) Our heaviness-representations allow for successful re-identification and discrimination of objects. (2) Our heaviness-representations are involved in mistaken, but arguably reasonable, inferences. (1) is explained by the reliability of heaviness-representations. Our heaviness-representations track objects' relational properties, and thus allow us

to re-identify objects when they trigger the same heaviness-representations and discriminate between objects when they don't. (2) is explained by misrepresentation. The inference is reasonable in that if an object really did have an intrinsic property of being heavy, then that should contribute to its being *generally* hard to lift, which should make it hard to lift on the moon. But the inference is unsuccessful because the dictionary doesn't really have an intrinsic property of heaviness.

Of course, there are other possible explanations of our heaviness-related behaviors and responses, but the example illustrates the type of explanation that is open to us if we allow for reliable misrepresentation: We can account for the successful execution of certain tasks that require reliability and the failure of certain other tasks that require veridicality. This is useful because reliability and veridicality contribute to successful behaviors and responses in different ways. Keeping them apart allows for explanations of patterns of behaviors and responses that would otherwise be foreclosed.

Again, I have not argued that such explanations are correct for any particular case. Instead, I am claiming that such explanations are potentially useful, and that we should not rule out such accounts of the functioning of any particular representation *in advance*. Whether or not such explanations are applicable should be decided by consideration of the particular cases in question, not on the basis of a metaphysical theory of mental representation.

6.2 The metaphysical reason

Another reason that a theory of mental representation should not preclude certain cases of reliable misrepresentation is that such preclusion leads to inappropriate metaphysical conclusions. A theory that prohibits reliable misrepresentation would force us to be realists about properties represented in nonsemanti-

cally successful conditions, where **realism** about a property P is the view that P is instantiated. If we have a representation R that represents P and occurs in nonsemantically successful conditions, then it follows that realism about P is true. This kind of inference would be inappropriate. A theory of representation is a theory that tells us how we get to represent. Realism about P is a view about the extra-mental world. It's just not the business of a theory of mental representation to settle the question of whether realism about P is true.[16]

Here is one way to see the problem: Determining whether realism is true of some item usually requires checking the world for the relevant item or for evidence of that item.[17] If we want to know whether to be realists about Bigfoot, we should look for Bigfoot or for relevant evidence for Bigfoot's existence (e.g. oversize footprints). But if tracking theories are correct, then in order to establish realism about a represented property P, we needn't check the world for evidence of instances of P. We can instead check ourselves for nonsemantically successful instances of the representation of P. Checking to see whether a token representation is nonsemantically successful doesn't even require specification of the represented property; we only need to check whether the representation is useful to our survival and flourishing, whether it occurred in circumstances of the same type as design conditions, or whether it bears a robust causal connection to anything—it doesn't matter what. From the facts that we have experiences of P and that the relevant conditions obtain, we can conclude that there are instances of P. While this method of determining whether realism about P is true involves checking the world, it involves checking the wrong parts of the

[16] I'm not claiming that it's always inappropriate for a theory to make predictions outside its domain. I suspect that which predictions of this sort are appropriate depends on our background theories, but this topic is far beyond the scope of this paper. In this section, I'm only claiming that it is inappropriate for the tracking theory to make predictions about realism about represented properties.

[17] There are exceptions, e.g. in the case of putatively necessarily existing items, such as numbers, where it might arguably be possible to determine whether they exist *a priori*. But the properties we represent in experience are clearly not all necessarily existing properties.

world: We don't have to check objects for P or traces of P. We don't even have to know what it would take for the world to count as having instances of P.

For example, if we want to know whether color realism is true, presumably we should check the surfaces of objects for properties that could plausibly qualify as colors. But if, say, we know that an optimal-functioning version of the tracking theory is correct, then all we would have to do is to (1) check our experiences to see if we ever represent colors and (2) check if representing colors is ever useful for us. First, these are the wrong kinds of facts from which to conclude that color realism is true. Second, we already know that (1) and (2) are true, so if we know that the optimal-functioning version of the tracking theory is correct, we can—*right now, without any further metaphysical discussion about what colors are, can be, or must be, and without any empirical examination of objects*—conclude that color realism is true! But we are in no such epistemic position. And that's why a theory of mental representation has to allow for clean cases of reliable misrepresentation. If it doesn't, it will have as a consequence that we are in these kinds of epistemic positions.

The situation is analogous to the following situation that is thought to be problematic for externalism: It is generally agreed that it is problematic if a theory of meaning or reference allows us to run the following argument:[18]

(P1) I have thoughts about water. (Introspective observation)

(P2) If I have thoughts about water, then water exists. (From theory of meaning or reference)

(C) Water exists.

If this type of argument is indeed unacceptable, then the following analogous argument should be likewise unacceptable:

[18]See Boghossian (1997) for discussion of this argument.

(P1′) I have experiences of redness. (Introspective observation)

(P1.5) My experiences of redness at least sometimes occur in nonsemantically successful conditions. (Uncontroversial empirical claim)

(P2′) If I have experiences of redness in nonsemantically successful conditions, then realism about redness is true. (From tracking theory)

(C′) Realism about redness is true.

From the fact that we have representations that are useful to us, occur in conditions of the same type as the conditions in which they occurred and were useful for our ancestors, or bear a robust causal relation to some property, we are forced to conclude that realism is true of the represented property. If we find the inference from representing water to being in a water world objectionable, we should likewise find this inference objectionable.

Premise (P1.5) does not have an analogue in the problematic externalist argument. Might this point to an asymmetry between the two cases that can be used to argue that the argument from color-experience to color realism is unproblematic? I think not. One way to see this is that we are in fact fairly certain that (P1.5) is true, and (P1′) is also true, but given our epistemic situation, it would be inappropriate *for us* to conclusively decide that colors are instantiated on the additional basis of a theory of mental representation. So (P1.5) is not the culprit. Another way to see why (P1.5) doesn't affect the general point is to repackage the argument to yield a slightly different argument:

(P1′) I have experiences of redness. (Introspective observation)

(P2′′) If I have experiences of redness, then either realism about redness is true or my experiences of redness are nonsemantically defective. (From tracking theory)

(C″) Either realism about redness is true or my experiences of redness are nonsemantically defective.

It would be inappropriate to conclude (C″) on the basis of (P1′) and one's theory of mental representation. (C″) is a disjunction of two empirical claims about extra-mental reality and particular mental states (one about colors, the other about particular mental representations and their uses for me or my ancestors, or their involvement in certain counterfactual dependencies). We should not be able to conclude that the disjunction is true from an experience of redness and a metaphysical theory of mental representation alone.[19]

All this is not to say that no facts about our experiences are relevant to the issue of realism about represented properties. Everything I have said so far is compatible with such facts being relevant to questions of realism in at least the following two ways: First, by examining the contents of our representations, we might gain insight onto what it would take for realism about a represented property to be true. For example, the contents of color-representations determine or at least constrain what would count as an instance of a color property, and so constrain how the world must be in order for color realism to be true. But this is just to say that in answering the question of whether the world is as it is represented by the mind, examining the contents of the mind can give us insight onto the mind side of the equation. Second, it's commonly thought that our color-experiences might offer us defeasible perceptual evidence that there are color instances in the world.

One might object that tracking theories are only known *a posteriori*, and it is not an objection to an *a posteriori* theory that it entails other *a posteriori* truths, such as (C″). But this objection rests on a misunderstanding of my

[19]It's not clear whether the tracking theory is supposed to be *a priori*. If it is, then the situation is even worse. We would be concluding that a disjunction of empirical claims is true on the basis of *a priori* truths and a seemingly unrelated experience. This shouldn't be possible. Boghossian (1997) does take his target theory to be *a priori*, but it's unclear whether the externalist need be committed to an *a priori* version of her view either.

argument: The trouble is not that tracking theories allow us to infer *a posteriori* truths from *a priori* truths, but rather that they allow us to make inferences that it seems we should not be able to make, whether or not any of the premises we use are *a posteriori*. Put otherwise, the problem is that from a tracking theory we can *a priori* infer conditionals such as "If (P1′) then (C′′)".

In conclusion, a metaphysical theory of mental representation should be compatible with the possibility of clean cases of reliable misrepresentation. Failure to be thus compatible inappropriately forces us to realism about certain represented properties.

6.3 Two responses

One might respond to my argument by agreeing that tracking theories cannot allow for clean cases of reliable misrepresentation, and maintaining that this is not a problem. I will consider two versions of this response here. These responses do not provide a direct reply to my arguments in §6.1 and §6.2, but if they are compelling, they might shift the balance of considerations in favor of the claim that the tracking theory need not allow for clean cases of reliable misrepresentation.

First, the tracking theorist might claim that although it cannot allow for *actual* cases of reliable misrepresentation, that's okay, because it can allow for *merely possible* cases. This is because the tracking theory is contingent. It need not be true in all possible worlds. In particular, it need not be true in possible worlds in which there are clean cases of reliably misrepresentation.

This position involves a concession to my arguments: The tracking theorist accepts the possibility of reliable misrepresentation of the relevant kind, and that the tracking theory must be rejected if such cases turn out to be actual. For example, the objector agrees that if it turns out that anti-realism about

colors, sweetness, attractiveness, morality, or other such properties is true, we will have to abandon the tracking theory.

This response also involves another concession, one that the tracking theorist really should not make. If there are merely possible clean cases of reliable misrepresentation, then the tracking theory cannot be taken to provide a sufficient condition for mental representation. This is because if clean cases of reliable misrepresentation are possible, then there are possible worlds in which representations bear robust, useful, and evolutionarily stable relations to something other than what they represent. But this requires giving up on the claim that the tracking theory offers sufficient conditions for mental representation, since if it did offer sufficient conditions, they would be met in these cases, and they would not count as cases of *mis*representation after all, but rather as cases of veridical representation of the tracked property.[20] And so, allowing for merely possible cases of reliable misrepresentation involves conceding that the tracking theory does not offer sufficient conditions for mental representation. But then in what sense is it a theory of mental representation?

The second response the tracking theorist might offer in favor of the claim that she need not allow for clean cases of reliable misrepresentation is this: For any putative case of reliable misrepresentation, we can reasonably deny that it is a case of reliable misrepresentation by denying that it represents what we think it represents and instead claiming that it represents what it tracks. For example, suppose the friend of reliable misrepresentation claims that color-representations reliably misrepresent: they represent simple color properties, which are

[20] One might claim that in these possible cases, representations have two contents because they meet two different sets of sufficient conditions for mental representation, those offered by a tracking theory and some other conditions. With respect to the tracking theory's content, they veridically represent, but with respect to the other content, they reliably misrepresent. Note that on this view, the tracking theory need not be contingent as claimed above. The problem is that since the resulting cases of reliable misrepresentation involve concomitant veridical representation of the tracked property, they do not look like *clean* cases of reliable misrepresentation.

not instantiated, but track complex surface reflectance properties. The tracking theorist might respond that color-representations do not in fact represent simple colors, but instead represent complex surface reflectance properties. She might further claim that the main reason we think that color-representations represent colors is intuition, and our intuitions should not be taken too seriously. In fact, she might claim, the tracking theory itself gives us a reason to think that what color-representations represent are surface reflectance properties. This kind of response can be offered for any putative case of reliable misrepresentation, thereby allowing the tracking theorist to deny all the problematic cases.

However, what a representation tracks is not the only evidence we have for determining what it represents. The friend of reliable misrepresentation need not rely on intuition alone, as the argument suggests, but can also make use of other sources of evidence, such as introspection, and behavioral and neurological data. This does not mean that our prior beliefs about mental representation are unrevisable, but only that we have some fairly theory-independent sources of evidence for determining what a representation represents. This evidence might agree with the verdicts of a tracking theory, but it might not. And this means that the tracking theorist does not have a fool-proof strategy for denying that any putative case of reliable misrepresentation of one thing is actually a case of veridical representation of something else. Each case must be considered separately, and it would be overly optimistic of the tracking theorist to assume that the balance of evidence would not favor clean reliable misrepresentation in any of the problematic actual or merely possible cases.

It is helpful to draw a comparison with the disjunction problem. The tracking theorist might respond to the disjunction problem by claiming that the problematic representations really do have disjunctive contents, since this is what is predicted by the tracking theory and any intuitions to the contrary are not to

be taken seriously. This response is not compelling because what a representation tracks is not the only evidence available for determining its content, and it would be overly optimistic for the tracking theorist to assume that the balance of evidence supports disjunctive contents in all possible problematic cases.[21]

6.4 Taking stock

In this section, I have argued that it is inappropriate for a theory of mental representation to rule out clean cases of reliable misrepresentation because this forecloses certain appealing psychological explanations and allows us to make inappropriate inferences about the non-psychological world. Thus, the difficulty that tracking theories face in allowing for reliable misrepresentation is a serious problem for them.

Of course, such considerations against tracking theories need to be weighed against other considerations we might have in favor of them, as well as the comparative plausibility of other views of mental representation. The balance of evidence might, for example, favor a view combining the tracking theory, the view on which clean cases of reliable misrepresentation are impossible, and realism about the required properties. I think this is unlikely, but it is beyond the scope of this paper to settle this. Instead, my present claim is that its

[21] This strategy for denying that problematic cases are genuine cases of reliable misrepresentation also risks trivializing the tracking theory. In order to qualify as a theory *of mental representation*, the tracking theory must allow that we have an antecedent grip on the phenomenon of mental representation. If we have no grip on the phenomenon of mental representation apart from that of tracking, then the tracking theory at best tells us that tracking is tracking, and perhaps that there are psychologically important tracking relations. But everyone can agree with that. That is not the interesting claim the tracking theorist wants to make. Compare: The mind-brain identity theory must allow that we have a grip on the notion of the mind independent of our grip on the notion of the brain in order for it to be an interesting theory *of the mind*, rather than a relatively uninteresting theory of the brain.

The problem is that our antecedent grip on mental representation involves a grip on the particular contents of particular states. This grip need not come from intuition alone, but can also come from introspection of our own mental states, and behavioral, psychological, and neurological data. This means that we have theory-independent criteria for determining what a representation represents, and that the tracking theorist cannot rely solely on the verdicts of her theory to determine the content of mental representations.

disallowing of clean cases of reliable misrepresentation is a strong consideration against tracking theories.

7 Conclusion

I have argued that tracking theories face difficulties in allowing for reliable misrepresentation. This is because reliable misrepresentations are *reliable*, and that means it is overwhelmingly likely that they are useful to us, that they were useful to our ancestors, and that they involve robust causal connections. In short, apart from not being veridical, reliable misrepresentations are very well-behaved.

Like illusion, hallucination, and occasional misrepresentation, reliable misrepresentation is a type of misrepresentation, and an adequate theory of mental representation should make room for it. Failure to allow for the possibility of reliable misrepresentation forecloses certain potentially useful kinds of psychological explanations and can also inappropriately force us to realism about certain (mis)represented properties. Making room for reliable misrepresentation, on the other hand, allows us to properly acknowledge two distinct ways in which mental representations can be successful: reliability and veridicality.[22]

References

Boghossian, P. A. (1997). What the externalist can know *A Priori*. *Proceedings of the Aristotelian Society*, 97(2):161–175.

Chalmers, D. J. (2006). Perception and the fall from eden. In Gendler, T. S. and Hawthorne, J., editors, *Perceptual Experience*, pages 49–125. Oxford University Press, Oxford.

[22]Thanks to David Bourget, David Chalmers, Gilbert Harman, Frank Jackson, James Martin, Daniel Stoljar, and an anonymous reviewer for helpful written comments on drafts of this paper, and to Joshua Knobe and Jack Woods for helpful discussion of the ideas in this paper.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press, Cambridge, MA.

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

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

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

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. A. (1994). *The Elm and the Expert*. MIT Press.

Holman, E. L. (2002). Color eliminativism and color experience. *Pacific Philosophical Quarterly*, 83(1):38–56.

Horgan, T. and Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In Chalmers, D. J., editor, *Philosophy of Mind: Classical and Contemporary Readings*, pages 520–533. Oxford University Press, Oxford.

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

Mendelovici, A. (2010). *Mental Representation and Closely Conflated Topics*. PhD thesis, Princeton University.

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. MIT Press.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86:281–297.

Pautz, A. (2006). Sensory awareness is not a wide physical relation: an empirical argument against externalist intentionalism. *Noûs*, 40(2):205–240.

Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press, Cambridge.