Why tracking theories should allow for clean cases of reliable misrepresentation

Angela Mendelovici

Disputatio 8 (42):57–92 (2016)
Penultimate draft

Abstract

Reliable misrepresentation is getting things wrong in the same way all the time. In Mendelovici 2013, I argue that tracking theories of mental representation cannot allow for certain kinds of reliable misrepresentation, and that this is a problem for those views. Artiga 2013 defends teleosemantics from this argument. He agrees with Mendelovici 2013 that teleosemantics cannot account for clean cases of reliable misrepresentation, but argues that this is not a problem for the views. This paper clarifies and improves the argument in Mendelovici 2013 and response to Artiga's arguments. Tracking theories, teleosemantics included, really do need to allow for clean cases of reliable misrepresentation.

Keywords: teleosemantics, representation, intentionality, misrepresentation, mental content

In "Reliable misrepresentation and tracking theories of mental representation" (Mendelovici 2013), I argue that a certain prominent class of theories of mental representation, tracking theories, have trouble allowing for what I call *clean* cases of reliable misrepresentation, and that this is a serious problem for them.

In "Teleosemantics and reliable misrepresentation," Marc Artiga (2013) provides an interesting defense of teleosemantics from this argument. Artiga agrees that teleosemantics cannot allow for the relevant kinds of clean cases of reliable misrepresentation, but argues that this is not a problem for the view. This paper clarifies the argument from reliable misrepresentation (section 1) and addresses Artiga's objections (2). If I am right, then tracking theories really do need to allow for clean cases of reliable misrepresentation.

1 The argument from reliable misrepresentation

The argument from reliable misrepresentation against tracking theories proceeds in two steps: Step One argues that tracking theories are incompatible with certain kinds of cases of reliable misrepresentation, *clean* cases of reliable misrepresentation. Step Two argues that this is a problem for these views. This section explains key notions and presents the two steps of the argument.

1.1 Key notions

Mental representation is the aboutness of mental states. A visual experience might represent that there is a cup on the table, a thought might represent that grass is green, and a desire might represent coffee, or that I obtain coffee.[1] I will assume that there are *mental representations*, which are are internal states that are the bearers of representational properties. What a mental representation represents is its *content*.

Tracking theories of mental representation are theories of mental representation that aim to account for mental representation in terms of causal or other

[1]My favored way of fixing on the phenomenon of mental representation is ostensive. See Mendelovici MS: ch 1 and Mendelovici 2010: ch. 2. See Mendelovici MS: ch 1 for a defense of the claim that other ways of picking out mental representation at least aim to include the ostensively defined phenomenon.

tracking relations between mental representations and properties, states of affairs, or other items. My argument is aimed at tracking theories in general, but Artiga is exclusively interested in defending teleosemantic tracking theories (see Millikan 1984, Papineau 1987), and in particular something close to Millikan's (1984, 1989) version of the theory. My main aim is to respond to Artiga's argument, so I will also focus on teleosemantics and a Millikan-esque version of the theory as well. However, much of the discussion applies to tracking theories more broadly.

Simplifying considerably, *teleosemantics* states that a representation R represents a content C if it is a normal condition for the systems that make use of R (R's consumers) to perform their proper functions that R corresponds to C. Let us upack this a bit: Representations have *producers* (or *senders*) that produce representations and *consumers* (or *receivers*) that respond to representations. Producers and consumers might be systems in a single organism, or they might occur in distinct organisms. An item or system's *proper function* is what items of its type did in its ancestors that resulted in their being selected for by evolution. For example, although a heart may do many things (pump blood, make noise, suffer certain kinds of blockages), it is its pumping blood that resulted in its being selected, and so its proper function is to pump blood. In order for something to perform its proper function in the way that it did that resulted in its selection, certain conditions have to be in place; these are *normal conditions*. The normal conditions for a heart performing its proper function include being part of an intact organism and receiving oxygenated blood. Something might perform its proper function while in conditions that are not normal (in other words, in *abnormal conditions*); in this case, its success at performing its proper function is in some sense accidental.

Like hearts, the consumers of representations have proper functions. It is a

normal condition for them to perform their proper functions that they correspond to certain states of affairs, which are their contents. Put otherwise, in order for representation consumers to perform their proper functions in the way that was selected for by evolution, a certain correspondence between representations and certain states of affairs had to be in place. Representations represent whatever this correspondence maps them onto.[2]

A characteristic example discussed by Millikan is that of beaver tail splashes (see, e.g., Millikan 1989). Beavers splash their tails to signal danger, which leads to other beavers taking cover. A tail spash at a location at a time is a representation, with the splashing beaver being the producer of the representation, and the onlooking beavers being the consumers of the representation. If a beaver flees and there is danger, the consumer succeeds in performing its proper function, which might be to avoid danger, in a normal way. If a beaver flees and there is no danger, the consumer does not succeed in performing its proper function in a normal way. The normal condition for the proper functioning of the representation's consumer is that there be a correspondence between the location and time of a tail splash and the location and time of danger. So, a tail splash at location l and at time t represents *there is danger at location* l *and time* t. While beaver tail splashes are not examples of *mental* representations, the same general principles apply to cases of mental representations.[3]

[2]A thorough development of Millikan's teleosemantics can be found in Millikan 1984. See also Millikan 1989 for an overview of some of the key ideas, and Shea 2004 for a lucid explanation of the view.

[3]Millikan, and I assume other advocates of the version of teleosemantics that Artiga aims to defend, rejects the view that *propositional representations*, representations representing entire putative states of affairs, are built up out of *subpropositional representations*, representations representing objects, properties, or other subpropositional contents (see, e.g., Millikan 1984: 107). Since normal conditions are entire states of affairs, the contents of mental representations are, in the first instance, propositional. (Millikan rejects a language of thought picture (Fodor 1975). See Rupert 1999 for discussion.) However, she takes propositional representations to have *variant aspects* and *invariant aspects*, aspects that do or do not vary, respectively, between different representations in the same system. For example, a beaver tail splash is a representation with a propositional form, representing the propositional content *there is danger at location* l *and time* t, for some l and t. The representation is part of a system of representations, a system of possible tail splashes, which have variant and invariant aspects,

1.2 Reliable misrepresentation

It is generally agreed that a theory of mental representation must allow for *mis*representation, representation that is false or inaccurate. In Mendelovici 2013: 423, I argue that just as we acknowledge hallucination, illusion, and occasional misrepresentation as kinds of misrepresentation, we should also recognize another kind of misrepresentation: reliable misrepresentation.

Intuitively, reliable misrepresentation is getting things wrong in the same way all the time (Mendelovici 2013: 422). Reliable misrepresentations are not just *wrong*; they are *systematically wrong*. Unlike hallucination and occasional misrepresentation, reliable misrepresentation is reliable; it involves misrepresenting *in the same way* all the time. Unlike illusion, reliable misrepresentation is not compatible with the overall veridicality of tokens of the representation in question; tokens of the representation in question are never veridical.

For example, suppose color anti-realism is true and nothing is colored. Then, color experiences reliably misrepresent. They represent that certain objects have certain colors, are always nonveridical, and occur reliably in the same sets of circumstances (e.g., ripe tomatoes tend to trigger representations of redness, a clear daytime sky tends to trigger representations of blueness). Color experiences get things wrong in the same way all the time. Similarly, suppose moral anti-realism is true and nothing is right or wrong. Then, our representations of rightness and wrongness reliably misrepresent. They represent that certain acts are right or wrong, are always nonveridical, and occur reliably in the same sets of circumstances (e.g., murders tend to be represented as wrong).

aspects that do and do not vary, respectively, between different representations in the system. The particular location and time can vary, so the aspects of the representation corresponding to location and time are variant aspects, while the representation of danger cannot vary and so it is an invariant aspect (e.g., beaver tail splashes cannot represent the location of elephants, rather than danger). These variant and invariant aspects play some of the same roles of subpropositional representations, such as that of accounting for productivity. (See Martínez 2013 for discussion.) So, for simplicity of exposition, I will take the term "representation" to cover these aspects.

While I think reliable misrepresentation is a natural phenomenon whose precise boundaries we might only hope to discover by empirically investigating its similarities and differences to other nearby phenomena, I will attempt to provide a more precise characterization of the phenomenon. Below is what a take to be an improvement over my characterization in Mendelovici 2013, discussion of which is relegated to a footnote.[4]

A representation R representing a property P reliably misrepresents for an organism O if and only if

(RM1) R represents P. (Representation)

(RM2) All of O's propositional representations that use R to ascribe P to something are false. (Nonveridicality)

(RM3) Ascriptions of P using R reliably co-occur with the presence of some type of state of affairs. (Reliability)

[4]In Mendelovici 2013 (p. 423), I characterized reliable misrepresentation as follows: An organism's representation of type R *reliably misrepresents* some property P if and only if

(RM1old) Some tokens of R are involved in attributive mental states that represent objects as having property P,

(RM2old) Most or all of the relevant objects do not have P,

(RM3old) Tokens of R do or would nonveridically represent objects as having P in the same types of circumstances on separate occasions.

The three clauses correspond to the three features of reliable misrepresentation: Representation, nonveridicality, and reliability. My proposed amended characterization impoves upon this characterization as follows:

First, as Artiga (2013) rightly points out, (RM2old) is weaker than I intended, since it allows the definition to include kinds of cases I do not want to call cases of reliable misrepresentation. (This has also been suggested to me by Frédéric-Ismaël Banville in conversation.) For example, suppose a representation R represents *danger*. Since, for many organisms, the cost of failing to flee when there is danger is greater than the cost of fleeing when there is no danger, it would be no surprise if such organisms would be set up so that most of R's tokenings will be false alarms. It might seem, then, that R would satisfy all the conditions for reliable misrepresentation. To avoid this, as Artiga correctly suggests, "most or all" in (RM2old) should be changed to "all" and calls the resulting notion that of "strong reliable misrepresentation." However, it is not clear to me that this case would satisfied the original definition, since it is not clear that the false alarms would occur *reliably*. Presumably all sorts of things could trigger the false alarms, so tokens of R would not tend to occur in the same circumstances. However, to avoid such worries, I have taken Artiga's friendly amendment on board.

Second, (RM3old) can be misread as stating that R is *nonveridical* in the same kinds of circumstances, rather than as stating that R occurs and is nonveridical in the same kinds of circumstances. My new characterization rules out this potential misreading.

In short, a mental representation reliably misrepresents for an organism just in case all its attributive uses in that organism are false and occur in the same kinds of circumstances.

There are two points worth noting about this characterization: First, a representation's reliably misrepresenting for one organism is compatible with the same representation not reliably misrepresenting for another organism. Suppose that color anti-realism is true and there are no colored objects on Earth, but there are colored objects on Colorful Earth, which is otherwise just like Earth. In this case, color representations reliably misrepresent for me, but not for my Colorful Earth twin.[5]

Second, a representation R's reliably misrepresenting is compatible with its occurrence in true propositional representations, so long as these propositional representations do not ascribe the property R represents to anything. For example, suppose again that color anti-realism is true and our color representations reliably misrepresent. Still, a thought that red is more similar to orange than it is to blue or that a particular object is *not* red might be true.

In the color case, reliable misrepresentation is systematic in a way that is not captured by my characterization: Not only does the representation of blue_{421} misrepresent whenever it is used attributively, but so too do related representations of other colors. The whole system of color representations reliably misrepresents. Whether we want to build this into our characterization of reliable misrepresentation is an interesting question. I take reliable misrepresentation to be a natural phenomenon, but it is unclear whether this kind of systematicity would be a feature of this phenomenon (one reason to think it might be is that it distinguishes reliable misrepresentation from certain kinds of

[5]I am characterizing reliable misrepresentation as a feature of types of representations that they have relative to organisms that have their tokens. But we might, instead, take reliable misrepresentation to be a feature of an organism's set of tokens of a representation or ability to token a representation.

illusions—see fn. 8). In any case, it seems that the prime examples of reliable misrepresentation are systematic in this way too.

1.3 Step One of the argument

We are now in a position to overview the argument from reliable misrepresentation against tracking theories. Step One of the argument argues that tracking theories of mental representation cannot allow for clean cases of reliable misrepresentation that are supposed to get their contents from tracking.[6]

Clean cases of reliable misrepresentation are cases of reliable misrepresentation that exhibit the likely features of reliable misrepresentation (Mendelovici 2013: 429). Since reliable misrepresentations are reliable, they are likely to be useful for their bearers, stable across generations, and such that they are tokened as a result of a robust causal connection, so clean cases of reliable misrepresentation have these features. However, this list of features is not meant to provide a complete characterization of clean cases of reliable misrepresentation. In getting a clearer idea of the notion, it is helpful to keep examples of putative clean cases in mind, such as that of perceptual color representations on the assumption of color anti-realism: Perceptual color representations form a system of representations, all of which misrepresent, occur reliably in the same circumstances on multiple occasions, but are useful for their bearers, stable across generations, and robustly causally connected to surface reflectance profiles. There is nothing "fluky" or contrived about this case. As we will soon see, it does not matter for the argument that the notion of a clean case of reliable misrepresentation is fuzzy, since, as long as there are cases of reliable misrepresentation that a tracking theory inappropriately rules out, the theory is in trouble. So this un-

[6]Note that this allows that there can be clean cases of reliable misrepresentation of representations that are not supposed to get their contents from tracking, e.g., representations that are supposed to get their content compositionally. I set aside such cases in what follows. Step One claims that some representations that are supposed to get their contents from tracking are clean cases of reliable misrepresentation.

derstanding of clean cases of reliable misrepresentation will suffice for present purposes.

While various tracking theories can allow for cases of reliable misrepresentation, the kinds of cases they can allow for are unclean in various ways. In the case of teleosemantics, reliable misrepresentation can occur in cases where our environment has changed from that of our ancestors. Suppose that our ancestors had an internal state, R, that co-occurred with the presence of items having property P and that this was helpful for our ancestors' survival and reproduction, so R represents the presence of P. Suppose, further, that P is part of a system of representations exhibiting variant and invariant aspects, and r is an invariant aspect corresponding to the property P. r would then represent P. But suppose now that our environment has changed since the time of our ancestors and there no longer are any Ps. Suppose also that r's misrepresentation is *reliable*; representations with aspect r occur regularly in some circumstances, say, in the presence of the property Q. We now have a case of reliable misrepresentation that is compatible with teleosemantics.

However, this setup tends to be *unstable*: selective pressure turns this case of reliable misrepresentation into a case of reliable *veridical* representation in subsequent generations. Since mere reliability tends to be useful for survival and reproduction (see Mendelovici 2013: 428-9), r's co-occurring with P is likely to be helpful to our survival and reproduction, and so, in subsequent generations, representations involving r will come to represent the instantiation of Q, and r itself with come to represent Q. Which descendants will be affected will depend on when selective pressure kicks in to preserve or change the proper function of the consumers of representations involving r, but as soon as such selection takes place, r comes to represent Q rather than P. The kind of reliable misrepresentation that teleosemantics can allow is unstable, giving way to reliable *veridical*

representation in one's descendants. Thus, although teleosemantics can allow for reliable misrepresentation, it cannot allow for *clean* cases of reliable misrepresentation. Similar kinds of arguments show that other kinds of tracking theories cannot allow for clean cases of reliable misrepresentation.[7]

[7] An anonymous reviewer suggests another way in which teleosemantics can allow for reliable misrepresentation. As noted above, it is possible for something to perform its proper function in abnormal conditions, or "by accident." This involves performing its proper function by doing something other than what its ancestors did that led to its being selected. To borrow an example from Millikan, a chameleon's pigment-changing mechanism might have the proper function of altering the chameleon's skin pattern to match what it sits on. A normal condition for its proper functioning is that it in fact matches what it sits on. But a chameleon's pigment-changing mechanism might still perform its proper function without matching what it sits on; it might cause the chameleon to stand in such stark contrast to its surroundings that a passerby feels sorry for it and moves it to a matching surface. The chameleon's pigment-changing mechanism performed its proper function, but not in the way that its ancestors did. It performed it "by accident."

Similarly, a representation's consumer can perform its proper function in abnormal conditions. If the environment changes such that normal conditions no longer ever obtain but abnormal conditions do sometimes obtain, a representation's consumer might continue to perform its proper function in these abnormal conditions. The suggestion is that the representation would then reliably misrepresent—all its instances would be false and occur in the same abnormal conditions. If there is no selective pressure against this setup, future generations might also come to have the same representation that reliably misrepresents.

However, in order for this to be a case of *reliable* misrepresentation, the representation would have to occur in the same abnormal conditions on multiple occasions. While it is possible that it does, it seems more likely that the representation would occur in all sorts of different conditions, and that, while sometimes its consumers will "get lucky" and perform their proper functions "by accident," in most cases they will not perform their proper functions at all. Consider again the lucky chameleon. Suppose the chameleon's environment changes and all the surfaces it can access are colors it cannot match. There are some surfaces it can match, but it can only access them by being placed there by a human being. The chameleon's pigment-changing mechanism can still perform its proper function, but only in abnormal conditions in which a human being feels pity for it and moves it to a matching surface. But, unless there happen to be sympathetic bystanders attending to it when required, its pigment-changing mechanism will sometimes change colors and not perform its proper function at all.

In order for the anonymous reviewer's case to be one of reliable misrepresentation, the representation in question will have to occur regularly in the abnormal conditions. But then if it occurs regularly in the abnormal conditions, and if its occurring regularly in these conditions helps its consumers perform their proper functions, then it is likely that natural selection will favor preserving the consumers' behaviors, and the abnormal conditions will become the normal conditions for proper functioning in future generations. Since representations represent the normal conditions for their consumers' proper functioning, the representation in future generations will no longer reliable misrepresent, but reliably *veridically* represent these new normal conditions. In short, the setup would be unstable.

Returning to the case of the chameleons, suppose that in the new environment, there are in fact sympathetic bystanders constantly attending to the chameleons. The pigment-changing mechanism is now useful to chameleons because it changes their skin to a pity-inducing color, which causes humans to move it to safety. But since the mechanism's behavior is useful, it is likely that selective pressure would help preserving it in future generations. As a result, the present generation's abnormal conditions for the proper functioning of the mechanism will become some future generations' *normal* conditions. The case is unstable.

Finally, even if the case of reliable misrepresentation were stable across generations, it would be unclean in other ways. For one, it requires consumers to perform their proper functions "by

The tracking theory's inability to allow for clean cases of reliable misrepresentation arises from the fact that tracking theories peg veridicality to various kinds of *nonsemantic success*, kinds of success distinct from veridicality. In the case of teleosemantics, a token representation is nonsemantically successful when it occurs in the same conditions that our ancestors found themselves in when the representation was useful for survival and reproduction. This does not mean that failure to be nonsemantically successful requires failure of a representation's consumers to perform their proper functions, but rather that the normal conditions for performing their proper functions do not obtain, so that if they do perform their proper functions, they do so "accidentally." Unfortunately, for representations that are supposed to get their contents directly from tracking, the conditions in which a representation is nonsemantically successful are of the same type as the conditions that fix the content of the representation, so it is impossible to misrepresent in nonsemantically successful conditions. This means that whenever misrepresentation occurs, something nonsemantic has to have gone wrong. Representations are never just nonveridical; they are nonveridical *and* unsuccessful in some other way as well. The problem with clean cases of reliable misrepresentation is that nothing nonsemantic has gone wrong. Clean cases of reliable misrepresentation are useful, stable, and the result of a robust causal relation. Apart from being false, they are perfectly well-behaved. Since tracking theories require misrepresentation to be accompanied by a nonsemantic defect, and since clean cases of reliable misrepresentations are nonsemantically successful, tracking theories cannot allow for clean cases of reliable misrepresen-

accident," which is very unlike the prime example of a clean case of reliable misrepresentation, that of the reliable misrepresentation of colors.

Relatedly, we should accept that there are other cases of *unclean* reliable misrepresentation apart from those described. For example, suppose, for some fluky reason, there is never nectar in a certain direction d of any bee hive. Then bee dances "stating" that there is nectar in direction d will misrepresent. Suppose further that there is an environmental condition that reliably causes bees to dance dances "stating" there is nectar in direction d. Then the bees' representation of d will reliably misrepresent, but this will not be a clean case of reliable misrepresentation.

tation.

1.4 Step Two of the argument

Step Two of the argument argues that it is a problem for tracking theories that they cannot allow for clean cases of reliable misrepresentation. This is because a theory of mental representation *should* allow for clean cases of reliable misrepresentation. Reliable misrepresentation is a kind of misrepresentation, just like hallucination, illusion, and occasional misrepresentation. Just as it would be inappropriate for a theory of mental representation to allow for only contrived cases of these other kinds of misrepresentation, it would be inappropriate for such a theory to allow for only unclean cases of reliable misrepresentation. In Mendelovici 2013: 436, I claim that it is clear from a pretheoretical perspective that a theory of mental representation should allow for clean cases of reliable misrepresentation, but I also offer two further considerations that support the claim.

The psychological consideration. The first consideration is that allowing for clean cases of reliable misrepresentation offers us the resources needed to make good sense of certain kinds of cases in which a representation helps us perform various tasks involving the discrimination and re-identification of objects, but is also involved in various mistaken inferences. For example, our representations of heaviness allow us to discriminate between objects of different weights and to re-identify objects based on how heavy they feel. However, our representations of heaviness lead to some mistaken inferences, such as that an object that is difficult to lift on Earth will also be difficult to lift on the moon. An appealing explanation of this pattern of reactions and inferences is that our perceptual representations of heaviness reliably misrepresent: they represent an intrinsic property of objects but reliably track a relational property holding be-

tween objects and other objects. It is an open empirical question whether this is the correct story of representations of heaviness; however, it is quite plausible that some representations might warrant such a treatment. This possibility should not be foreclosed on the basis of a theory of mental representation.

The metaphysical consideration. The second consideration in favor of allowing for clean cases of reliable misrepresentation is that failure to do so leads to surprising and unwarranted metaphysical conclusions, such as realism about various represented properties. The tracking theory licenses inferences from the fact that we represent P in nonsemantically successful conditions to realism about P, where *realism* about P is the view that P is instantiated. For example, if we represent colors in nonsemantically successful conditions, then the tracking theory allows us to conclude that color realism is true. In other words, the tracking theory entails the following conditional for any content P that we represent by tracking:

(REAL) If P is represented in nonsemantically successful conditions, then realism about P is true.

(REAL) is a consequence of any theory of mental representation that doesn't allow for misrepresentation in nonsemantically successful conditions. The problem is that (REAL) is false. Accepting it is incompatible with the normal ways in which we think we should settle questions of realism. In cases where the existence or non-existence of P is contingent, we normally think that in order to decide on realism about P, we should figure out what would count as an P and then check the world for evidence of P. In cases in which the existence of P is supposed to be necessary or a priori, different methods apply; for instance, we might consider whether P plays certain theoretical roles that cannot be played by anything else. But (REAL) allows us to bypass these usual methods. We

can tell that the antecedent of (REAL) is satisfied without figuring out what would count as a P is and checking the world for signs of P, or employing other normal methods for finding out about P. All we have to do is check ourselves for experiences of P and check to see if those experiences are nonsemantically successful. But this makes establishing realism too easy.

One way to summarize the two considerations is this: Tracking theories inappropriately prejudge certain empirical questions, questions concerning certain psychological and metaphysical facts. This is inappropriate in part because, despite being touted as "naturalistic" and in line with scientific world views, tracking theories are often established without investigating the psychological and metaphysical facts they prejudge.

Above, I noted that there is some fuzziness in the notion of a *clean* case of reliable misrepresentation. But we can now see why this does not affect the argument: The problem is not so much that tracking theories cannot allow *any* instances of a general category of misrepresentation, but that there are such instances that they should not disallow.

2 Response to objections

This section overviews and responds to various objections recently made in Artiga 2013. Artiga mainly agrees with Step One of my argument.[8] His dis-

[8] Artiga presents the case of the Ebbinghaus illusion as an example of reliable misrepresentation that teleosemantics can account for, although he does not claim this is a clean case of reliable misrepresentation, and so he does not take this to challenge Step One of my argument. While I agree with Artiga that teleosemantics can account for this case, it is not clear to me that it is a case of reliable misrepresentation at all. Artiga writes:

> [I]n the best-known version of the Ebbinghaus illusion, two circles of identical size are placed near to each other and one is surrounded by large circles while the other is surrounded by small circles; the first central circle then appears smaller than the second central circle. In this case, the selection and existence of a mechanism producing representations of size is explained by the fact that most of the time it produces the right representations. Nevertheless, there is a representation type R (the state that misrepresents the size of [the] inner circle

agreements are with Step Two.

> in the Ebbinghaus scenario) that reliably and systematically misrepresents a certain configuration. This state R, which reliably misrepresents an inexistent size of certain circles, is a by-product of the representational system that has earned its keep in evolution. This is a sort of case involving strong reliable misrepresentation that can be perfectly accommodated within teleosemantics. (2013: 270-1, footnote suppressed)

There are two ways of understanding what Artiga takes R to represent. On one interpretation, R represents "the size of [the] inner circle," (271) but on another interpretation, R represents "a certain configuration" (271). On the first reading, R is not a case of reliable misrepresentation, since many occurrence of the representation of the particular size in question are veridical. For example, a circle outside the context of an Ebbinghaus illusion might trigger a representation of a circle as being that size. On the second interpretation, R is a representation of the entire configuration of the inner and outer circles. Now, R should not be a propositional representation to the effect that there is such-and-such arrangement of circles, since the characterization of reliable misrepresentation only applies to representations of *properties*. Instead, it should represent a property that captures the arrangement of circles. For example, perhaps R represents a property attributed to space, that of *containing such and such an arrangement of circles.* Or perhaps R represents a property attributed to sets of objects, that of *being circles arranged in such and such way.* Then, one might argue, R, which represents some such property, misrepresents every time it is tokened in a particular individual and satisfies the other conditions on reliable misrepresentation. However, I don't think it's clear that we have representations that represent such properties. While it is plausible that the representational states we have while viewing the Ebbinghaus illusion involve a representation representing the size of the inner circle, representations representing the size of the outer circles, and a representation that there are such and such circles standing in such and such relations before us, it is not clear that we *also* count as having a representation corresponding to the composite properties discussed above. This would require that the representations representing the sizes of the inner circle and the outer circles come together in a way that qualifies them as a further representation without thereby representing an entire proposition. This would be analogous to a beaver's tail splash representing *danger at location l and time t* involving a representation or aspect representing *l and t.* It is an empirical question whether we have such a representation (or aspect of a representation), but I am doubtful.

A better example of a persistent illusion that might be a case of reliable misrepresentation is the experience of the Penrose triangle, since in this case it is more plausible that we have a single representation representing an (impossible) shape property. However, this kind of case is very different from paradigm cases of reliable misrepresentation, and might, at best, point to an inadequacy of the characterization of reliable misrepresentation. In paradigm cases of reliable misrepresentation, such as the case of color, misrepresentation is not merely a special case of a system of representations that might be otherwise generally veridical, but is instead a feature of the core cases of the system of representations. Additionally, if reliable misrepresentation is to be contrasted with illusions, then illusions should not generally qualify as cases of reliable misrepresentation.

A further condition that we might add to the characterization of reliable misrepresentation to rule out persistent illusions is the requirement that all representations of properties in a system of representations satisfy the first three conditions of reliable misrepresentation (see p. 7). Alternatively, we could say that this is not a *clean* case of reliable misrepresentation, perhaps on the basis of the fact that it is an isolated case of misrepresentation in an otherwise generally veridical system.

In any case, it does not really matter if persistent illusions are clean cases of reliable misrepresentation, since Artiga and I agree that, regardless of how they are classified, they are cases that teleosemantics can accommodate, and also that there are other cases that teleosemantics cannot accommodate. Our disagreement is over whether it is problematic that teleosemantics cannot accommodate the latter cases, not over whether some other cases that it can accommodate should be classified as the same type of case on some method of classification.

That's just what the theory says! Artiga writes:

> I agree [that the possibility of stable cases of reliable misrepresentation] is certainly ruled out by teleosemantics; according to the theory, R represents Q iff Q causally explains the selection of the mechanism producing R. However, that does not seem to be an unwelcome result, but just a different way of stating the theory. If R represents whatever feature explains its selection, it cannot happen that a feature explains its selection and it is not represented by R. Why should that be a problem? (2013: 273)

I'm not sure if the teleosemanticist needs to say that represented properties causally explain the selection of the mechanisms producing the representation in question, but I do agree that she is committed to represented properties playing some causal role in selecting for a representation's producing or consuming systems. I'm also not sure that teleosemantics needs to be committed to providing necessary and sufficient conditions for mental representation, rather than just sufficient conditions that apply to us. But these quibbles are inconsequential to Artiga's point, which is just that the impossibility of clean cases of reliable misrepresentation is part of what teleosemantics says. It's not an objection to a theory that it says what it says.

Artiga also puts the point in a different way:

> Every theory, teleosemantics a fortiori, is such that whatever meets the sufficient conditions for being an F according to a theory is an F according to the theory... In teleosemantics, those sufficient conditions involve a process of reliability and stability for a period sufficient for selection of the sender-receiver configuration. Consequently, it is certainly true that teleosemantics rules out a case in which Q is the property that accounts for the selection of R and R

does not represent Q... (2013: 273)

Although Artiga does not put things in quite this way, one way to understand his complaint is as being that the objectionable consequence of teleosemantics (its inability to allow for clean cases of reliable misrepresentation) is an immediate consequence of the theory, and objecting to a theory on the basis of its immediate consequences is question-begging, since it would not convince anyone who accepts the theory.

It is true that the failure to allow for clean cases of reliable misrepresentation is a consequence of teleosemantics. It is also true that it is possible to reformulate teleosemantics and the conditions in which the relevant kinds of clean cases of reliable misrepresentation occur such that this consequence can be recognized fairly immediately. Perhaps, on some ways of understanding teleosemantics and on some ways of understanding the allegedly problematic consequence, teleosemantics and the problematic consequence are equivalent. However, this is neither here nor there. A consequence of a theory is a consequence of a theory, no matter how immediate or how obvious it is on various formulations of the theory and the consequence. The question that should concern us is whether the consequence is acceptable. In Step Two of my argument, I offer specific reasons for thinking that the consequence is unacceptable. One reason is that it precludes certain kinds of explanations of certain patterns of reactions and inferences. Another reason is that it warrants certain inappropriate metaphysical conclusions. Another, fairly flat-footed, but I think compelling, reason is that reliable misrepresentation is a type of misrepresentation, just like occasional misrepresentation, hallucination, and illusion, and insofar as a theory should allow for uncontrived cases of any kind of misrepresentation, it should allow for clean cases of reliable misrepresentation. The mere fact that some claim is a consequence, or even a direct consequence, of a theory does not make

it immune from criticism.

Objection to the psychological consideration. Artiga (2013: 274, fn. 5) addresses the psychological reason for allowing for clean cases of reliable misrepresentation. The claim from Mendelovici 2013 under dispute is that reliable misrepresentation might be the best explanation of certain patterns of reactions and inferences, so we should not rule out the possibility of this sort of explanation on the basis of a metaphysical theory of mental representation. Artiga objects that, since certain (unclean) cases of reliable misrepresentation are allowed by teleosemantics, this form of explanation is not in fact ruled out.

All this is true. However, the relevant kind of explanation is only allowed in cases involving instability, that is, cases in which the relevant patterns do not persist over enough generations for there to be the kind of selective pressure that could confer a change in content to the representation in question. In stable cases, cases in which the relevant patterns persist over generations in the relevant way, the patterns cannot be explained in the way I describe. Since the reasons for wanting to allow for the kinds of explanations I describe have nothing to do with instability, it is inappropriate to restrict their application to unstable cases.

The allegedly inappropriate metaphysical consequences are just fine if we assume teleosemantics. Artiga takes issue with the metaphysical consideration in favor of allowing for reliable misrepresentation. Recall that my worry is that tracking theories entail (REAL).

(REAL) P is represented in nonsemantically successful conditions, then realism about P is true. (From a tracking theory)

The problem is that (REAL) licenses inappropriate inferences from claims that we represent P in nonsemantically successful conditions to realism about P,

which is in tension with how we generally think we can and cannot settle questions of realism.

Artiga agrees that the tracking theory licenses such inferences, but disagrees that this is a problem if we assume teleosemantics.

> In teleosemantics 'P is represented in non-semantically successful conditions' should be spell[ed] out as the claim that P was the property that accounted for the selection of the sender-receiver system. Hence, the problem should be cashed out as follows: if teleosemantics is right and P is the property that accounts for the existence and selection for the system, then one is committed to the (past) existence of P. Again, this conditional seems true, but also entirely plausible. If a property P accounted for the existence of the representational system, P must have been instantiated somewhere. (2013: 275, footnote suppressed)

Artiga offers further reasons for thinking that it is okay to assume that represented properties were instantiated, which I will turn to shortly. But let us consider this argument first. Artiga is claiming that, from the perspective of the teleosemanticist, the allegedly unwarranted inference that is licensed by the tracking theory is in fact entirely warranted. This is because the teleosemanticist understands the antecedent of (REAL) as equivalent to "P was the property that accounted for the selection of my sender-receiver system," which makes (REAL) equivalent to (REAL$'$).[9]

(REAL$'$) If P was the property that accounted for the selection of my sender-receiver system, then realism about P is true.[10]

[9] The teleosemanticist has room to deny that "P is represented in non-semantically successful conditions" is equivalent to "P was the property that accounted for the selection of the sender-receiver system" if she claims to only provide sufficient conditions for mental representation and not necessary conditions or if she weakens the modal strength of her theory, but I set this aside for now and assume that the teleosemanticist accepts this equivalence.

[10] For present purposes, we can assume that the past existence of P is sufficient for realism

While (REAL) seems to be in tension with our ordinary ways of finding out whether realism is true, (REAL′) does not. However, this does nothing to show that (REAL) is unproblematic. This is because (REAL′) does not have the objectionable features of (REAL); it does not commit itself to a questionable connection between what we represent in nonsemantically successful conditions and realism.

The situation is analogous to the following: Suppose the *I'm Psychic Theory* claims that I imagine something iff it's true. This theory entails (PSYCH):

(PSYCH) If I imagine that P, then P.

One might object to (PSYCH) in various ways; for one, it is incompatible with how we normally think we can come to know about future events. Normally, we predict the future on the basis of present events and theories about what kinds of events they might give rise to. But (PSYCH) allows me to bypass such tedious methods and predict the future solely based on what I imagine, which may have no causal connection to the alleged future events I predict. According to the I'm Psychic Theory, however, I imagine something iff it's true. So "I imagine that P" is equivalent to "P," and (PSYCH) is equivalent to (PSYCH′):

(PSYCH′) If P, then P.

(PSYCH′) is clearly true and has no objectionable epistemological consequences. But this does nothing to show that (PSYCH) is true and similarly unobjectionable. (REAL) and (PSYCH) serve as bridge premises linking claims about representation and imagination, on the one hand, and claims about realism and truth, on the other. But (REAL′) and (PSYCH′) are not bridge premises; they operate only on one side of the relevant chasm, the realism/truth side. So,

about P. However, if the antecedent of (REAL′) is suitably cashed out so as to fully capture the commitments of the antecedents of (REAL) assuming teleosemantics, it would also state or entail that my sender-receiver system is stable in the relevant way and that I have tokens of the internal states tracking P. This would entail the present instantiation of P.

(REAL′) and (PSYCH′) do not have the objectionable features of (REAL) and (PSYCH). We can see this clearly by noting that (REAL′) makes no mention of mental representation and (PSYCH′) makes no mention of imagination.

Innately represented properties must have been instantiated in the past. In his paper, Artiga argues that the following principle is true:

(PAST) If a property P is innately represented, P was instantiated in the past.

The argument for (PAST) proceeds by attempting to show that (PAST) receives widespread and legitimate endorsement by scientists and philosophers. If it is unobjectionable to assume (PAST), this might motivate the claim that (REAL) is unobjectionable independently of Artiga's previous argument. The problem, however, is that there is little reason to think that (PAST) receives widespread and legitimate endorsement.

As an example of the endorsement of (PAST) in philosophy, Artiga cites an objection to Fodor's (1975) concept nativism:

> Some people have suggested that human concepts like CARBURETOR or TELEVISION cannot be innate because if they were, we would have to accept that there were carburetors and televisions at the time our ancestors evolved (Sterelny 1989; Prinz 2002: 229). (p. 275)

However, this is not in fact the argument that Sterelny and Prinz make. They argue that concepts like CARBURETOR are not innate as follows: Innate representations have to be selected for. In order for a representation to be selected for, it has to have been useful to our ancestors. But concepts like CARBURETOR would not have been useful to our ancestors. So, they could not have been selected for. So, they are not innate.[11] Sterelny and Prinz's line of argument

[11] Prinz writes: "If prevailing theories of evolution are true, innate representational resources

does not assume (PAST).

Sterelny and Prinz's argument is more compelling than Artiga's suggested reconstrual. The problem with thinking that CARBURETOR was selected for is not that there were no carburetors in our ancestors' environment, but that the concept CARBURETOR would have been useless for our ancestors. Now, perhaps part of the reason why the concept would have been useless for our ancestors is that there were no carburetors, but this is neither here nor there. The concept could have been useless even if there were carburetors (indeed, it is arguably useless to many people who have it today), and the concept could have been useful even if there were no carburetors (Mendelovici 2013 offers examples of such cases). Usefulness and accuracy can come apart, and it's usefulness that matters for selection, not accuracy (except insofar as accuracy confers usefulness).

As an example of the endorsement of (PAST) in science, Artiga presents the debate over the innateness of a fear of spiders. A consideration against the claim that a fear of spiders is innate is that only a small percentage of spiders are poisonous, and so a fear of spiders would not confer a significant selective advantage and would not have been selected for. Thus, it is not innate. Artiga takes this line of argument to assume (PAST). But it clearly does not for the same reason that Sterelny and Prinz's arguments do not assume (PAST). What is required for a fear of spiders to be selected for is that it is useful to our ancestors, not that it represents accurately. A representation's usefulness and accuracy can come apart, allowing for useful inaccurate representations and accurate useless representations.

---

must be either selected for or generated as an accidental by-product of things that were selected for. A concept like SPATULA could not have been selected for, because it would have conferred no survival advantage in the environments in which humans evolved." (2002: 229)

Similarly, Sterelny writes: "innate concepts require a selective explanation; an explanation showing *that very concept* conferred a reproductive advantage on our ancestors." (1989: 123, emphasis in original)

In summary, Artiga's arguments for the claim that (PAST) receives widespread and legitimate endorsement among philosophers and scientists are unsuccessful. As I've argued, considered endorsement of (PAST) is neither widespread nor legitimate.[12]

We don't know what we represent through introspection. In Mendelovici 2013, I provide a specific example of the kind of argument (REAL) licenses:

(P1) I represent *redness*. (Introspective observation)

(P2) My representations of redness occur in nonsemantically successful conditions. (Uncontroversial empirical assumption)

(REAL) If P is represented in nonsemantically successful conditions, then realism about P is true. (From a tracking theory)

(C) Therefore, realism about redness is true.

Importantly, both (P1) and (P2) can be known without examining the world for traces of redness. (P1) can be known through introspection, and (P2) can be known through examination of the usefulness, stability, and robustness of our representation. We can assume that the tracking theorist accepts that we can also know that a tracking theory is true, and hence that (REAL) is true, without

[12] Although Artiga's paper suggests that the principle he aims to defend is (PAST), in conversation, he has suggested a weaker principle:

(PAST-weak) If a property P is innately represented, P is likely to have been instantiated in the past.

The claim that (PAST-weak) receives legitimate and widespread endorsement is more plausible. However, this wouldn't help the teleosemanticist, since (PAST-weak) is too weak to legitimize (REAL). (REAL) does not claim that realism about properties represented in nonsemantically successful conditions is *likely* to be true, but rather that realism about such properties *is* true.

In summary, neither appeal to (PAST) or (PAST-weak) succeed at legitimizing teleosemantics' commitment of (REAL).

examining the world for redness.[13] The worry, then, is this: The tracking theory allows us to move from an introspective observation and an uncontroversial empirical claim to realism about redness, without even requiring us to examine the world for redness. This is incompatible with our ordinary ways of settling questions of realism.

Artiga agrees that teleosemantics allows one to move from (P1) and (P2) to (C), but maintains that this is unobjectionable. He suggests that I think that (P1) and (P2) are a priori and that this is precisely what bothers me.[14] But this is not what I think and this is not what bothers me. Nowhere do I say that (P2) or (P2) are a priori.[15] (P1) is a posteriori because it is known through experience, even though the relevant experience is introspection.[16] (P2)

[13]David Bourget has suggested to me that the tracking theorist could respond that we are not in a position to know that (REAL) is true without first determining that realism about properties represented in nonsemantically successful conditions is true. If this is right, then it is not problematic that (REAL) licenses a move from (P1), (P2) to (C), since properly justifying (REAL) requires already knowing the truth of (C). This is an interesting line of response on behalf of the tracking theorist, though, as Bourget points out, it comes with unfortunate consequences. It would mean that tracking theories have been accepted on inadequate evidence all along. In order to properly justify a tracking theory, it is not enough to show that it adequately accounts for cases of beavers, frogs, magnetotactic bacteria, and even some cases of beliefs and desires; the tracking theorist must also argue for realism about colors and other represented properties.

[14]Artiga writes: "Since P1 and P2 seem to be priori and C is clearly a posteriori, if we accept that P1-[(REAL)] entail C, we will be entitled to conclude a substantive and a posteriori claim about the world (color realism) from certain a priori claims and teleosemantics. I think this is precisely what worries Mendelovici. . . " (2013: 277)

[15]In Mendelovici 2013, I write: "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." (440)

Artiga acknowledges that I deny that my worry concerns moving from a priori premises to a posteriori conclusions, but interprets the following passage from my paper as nonetheless supporting his interpretation: "But if tracking theories are correct, then in order to establish realism about 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." (2013: 437-8) It might sound like my claim that, on the tracking theory, "we needn't check the world for evidence of instances of P" in order to draw realist conclusions about P means that the tracking theory allows us to draw realist conclusions about P without checking the world at all, i.e., a priori. But being able to draw realist conclusions about P without checking the world *for* P is compatible with having to check the world *for something* in order to drawing realist conclusions about P. What we have to check the world for is the truth of (P1) and (P2), which, as I claim in my paper, does not require checking the world for P.

[16]Like Millikan, Artiga takes introspective knowledge of our own mental states to count as a priori, so Artiga's disagreement with me over whether such introspective knowledge is a

is a posteriori because it is also known through experience; the fact that it is uncontroversial does not make it a priori. The problem is not that a posteriori conclusions can be drawn from a priori premises, but that questions of realism can be settled on the basis of a theory of mental representation and a posteriori but fairly innocuous and easily justified claims like (P1) and (P2).

Nevertheless, Artiga's response to his apparent misconstrual of my argument can also be offered as a response to the argument I actually make. His response is that once we appreciate the kind of empirical examination required to establish (P1) and (P2), it should not be surprising or objectionable that a theory of mental representation allows us to conclude from them that color realism is true:

> Teleosemantics is an externalist theory about content, so P1 and P2 are a posteriori claims through and through. What kind of property I am representing with a red experience and what kind of situations are nonsemantically successful conditions (i.e. what sort of situations accounted for the selection of the mechanism) are hard empirical questions that should be resolved by science. Consequently, even if teleosemantics is right, a considerable amount of empirical knowledge must be gathered before anything like C can be established. (2013: 278)

In order for this kind of response to successfully respond to the argument I actually made, it should show that establishing the conjunction of (P1) and (P2) requires establishing that color properties are instantiated through normal, presumably empirical, methods. We should be clear that everyone should grant that there are ways of establishing (P1) and (P2) that would proceed via establishing that color properties are instantiated—for instances, we could establish

priori is likely merely terminological.

that (P2) is true by establishing that in nonsemantically successful conditions, color properties are instantiated and then establishing that we find ourselves in such nonsemantically successful conditions. But the question is not whether there are ways of establishing either (P1) or (P2) that proceed via first establishing color realism, but rather whether there are ways of establishing (P1) and (P2) that *do not* require first establishing color realism. I claimed that there are.

Let us consider both premises separately to see if there is reason to think that establishing either premise requires establishing color realism. For teleosemantics, a representation's nonsemantically successful conditions are what we might call its *design conditions*, the type of conditions in which the representation's occurrence in our ancestors helped them survive and reproduce. Now, one way of establishing that we represent colors in design conditions is by first finding out precisely *what* the relevant design conditions are and then establishing that we represent colors in those conditions. This way of establishing (P2) would indeed involve establishing color realism prior to accepting (P2), since design conditions would have to involve the instantiation of color properties. (This is the way Artiga seems to have in mind when he says in the above quotation that "what kind of situations are nonsemantically successful conditions (i.e. what sort of situations accounted for the selection of the mechanism)" is an empirical question to be settled by science.) However, another way of establishing that we represent colors in design conditions is by first establishing that representing colors aids us in our survival and reproduction in reliable and systematic ways and then inferring from this that being in the same kinds of states helped our ancestors in the same ways. Unlike the way Artiga seems to have in mind, this perfectly good way of establishing (P2) does not require that we first establish color realism.

Let us now turn to (P1). Does establishing (P1) require first establishing that color realism is true? There are several ways of establishing what we represent without first having to establish that what we represent is true, exists, or is instantiated. Perhaps the most obvious way comes from introspection: In some cases, introspection affords us access to the contents of our representational states; it tells us which contents we represent. The case of conscious representational states, such as conscious thoughts or perceptual experiences representing redness, seem to be good candidates for states that allow for this kind of introspective access. It is important to note that none of this requires that introspection can reveal the content of *all* mental representations (perhaps we have non-conscious representations that are introspectively inaccessible), or that introspection is an infallible guide to content. All that is required is that, in some cases, introspection provides us fairly good evidence that we represent certain contents. Assuming the case of color representation is one of those cases, then this is enough to establish (P1), the claim that we represent colors.[17]

Artiga (and Millikan) reject the claim that introspection provides special access to representational states, so they would not accept this way of establishing (P1).[18] I think this position is overly skeptical about introspection. That I rep-

[17]This also does not require that introspection can reveal the metaphysical nature of mental representational states or their contents. See Mendelovici forthcoming, MS: ch. 1, where I argue that introspection can at least sometimes tell us *which* contents we represent without revealing to us their metaphysical nature (e.g., whether they are sets of possible worlds or structured propositions) or the metaphysical nature of mental representation in general (e.g., whether it is a tracking relation, a relation to abstract entities, a relation to sense data, or a non-relational state of subjects).

[18]As an anonymous reviewer has pointed out, a central part of Millikan's view is the denial of what she calls "meaning rationalism," which includes the view that introspection and intuition provide insight into the contents of our representational states (see especially Millikan 1984: 91-2 and 326-7). However, meaning rationalism involves a commitment to the infallibility of introspection, which I do not need in order to make my argument.

Millikan's own view of self-knowledge is, roughly, that knowledge of what our concepts represent is a matter of our (fallible) abilities to tell that two thoughts represent the same content (Millikan 2000: chs. 10 and 13). "Knowing what I am thinking of is being capable of coidentifying ... various of my thoughts with other thoughts of the same. It is being able to distinguish thinking of a thing again from thinking of a different thing." (Millikan 2000: 184) (See also Shea 2002 for an overview.) Millikan does claim that this picture is only "[t]he closest thing that actually makes some sense ... to the yearned-for ideal of comparison of a thought with its object bare within thought itself" (2000: 184), but it is not clear that it

resent colors is immediately obvious to me. I need not consider my dispositions to make inferences or behave, or my social or physical environment in order to know that I represent colors. That we represent colors is a datum, one that a theory of mental representation has to explain. Further, one might argue, it is through introspection that we get a grip on mental representation in the first place, and this way of getting a grip on mental representation automatically affords us some pre-theoretic access to mental representational states. But this is not the place to argue for these claims.[19] Instead, let me turn to another way in which we can find out about the contents of our representational states without introspecting upon them: by observing their psychological roles.[20]

Mental representational states play certain psychological roles, including roles in inference, behavior, and the formation of higher-order thoughts about

comes close at all, since it only seems to deliver knowledge that two concepts represent the same unknown thing, and knowing that two concepts represent the same content does not help you know what that content is if you have no prior access to either thought's content. The kind of knowledge we obtain is analogous to the knowledge we obtain by learning that two words whose meaning we do not know are synonymous.

[19] See Mendelovici 2010: ch. 2, MS: ch 1.

[20] Artiga's rejection of introspection involves conceding that externalist views like teleosemantics are in tension with introspective self-knowledge. He suggests that this tension shows there is nothing new about my argument, since we already knew that externalist theories like teleosemantics are in tension with introspective self-knowledge (2013: 278). Of course, that externalism is incompatible with introspective self-knowledge is not what my argument intends to show. My argument assumes a premise that is accepted by most participants on the debates concerning introspective self-knowledge, which is that there is such a thing as introspective self-knowledge, and attempts to show that together with other assumptions, this gives rise to consequences concerning color realism and other forms of realism. My overall argument is meant to show that clean cases of reliable misrepresentation are incompatible with tracking theories of mental representation and that this is a problem for them. So, my argument does not boil down to pointing out the tension between externalism and introspective self-knowledge. More generally, when one responds to an argument by biting a bullet, one cannot conclude that the aim of the argument was to establish the bullet.

Artiga also suggests that if the worry boils down to a worry about the compatibility of tracking theories with introspective self-knowledge, then it is a problem for any externalist theory, not just teleosemantics or tracking theories more generally, so "a defense will have to come from externalism, rather than from teleosemantics." (2013: 278) Of course, a problem for everyone is not a problem for no one, so, to the extent to which accommodating introspective self-knowledge is a problem for externalism, it is a problem for tracking theories, including teleosemantics. In any case, tracking theories are the main contenders for externalist theories of meaning, so even if my argument did boil down to pointing out the tension between externalism and self-knowledge, and even if a problem for all versions of a theory is not a problem for any specific version of that theory, the worry would still be a fair one to raise against my target, which is tracking theories in general.

them. For example, from representing that some object has a particular color, we are likely to represent that it doesn't have certain other colors, that it is a visible object, and that it is similar to or different from other objects. We might utter certain words, like "This is red," or approach or move away from it. We might form a higher-order thought with the content *I am thinking about a red tomato*. From these and other facts like them, we can hone in on the content of color representations without requiring us to know whether colors are in fact instantiated. Again, this method does not need to be foolproof in order to provide sufficient evidence for (P1).[21]

More generally, the tracking theorist should accept that there is a theory-independent way of finding out what we represent, a way that doesn't require first finding out what we track on their favored tracking theory. As long as this way does not require establishing realism about represented properties, we have a way of establishing (P1) independent of establishing realism about redness, and we have an argument for color realism that bypasses the normal ways of finding out whether realism is true. I will return to this point shortly.

We don't know *that* we represent through introspection. Artiga argues that the rejection of color realism is compatible with teleosemantics and the validity of the argument from (P1), (P2), and (REAL) to (C):

> [I]f we assume teleosemantics and grant everything I accepted in this paper (including the inference from P1-[REAL] to C), is teleoseman-

[21] One might object that knowing that we make inferences with a certain content or have higher-order thoughts requires introspection of that content, so although this method allows us to avoid introspection upon the representational states we want to know about, it does not avoid introspection entirely. I agree that the most natural ways of finding one whether you are making a particular inference is to introspect, and the most natural way of finding out whether someone else is making an inference with a particular content is to ask them whether they are, which will prompt them to introspect and then report on what they find. This dependence on introspection of the most natural way of finding out about what inferences we make illustrates the far-reaching consequences of rejecting introspection. But presumably the skeptic about introspection will allow that there are other ways of finding out about what inferences we make, perhaps through our behaviors. Something similar can be said for how we know about our higher-order states.

> tics still compatible with color eliminativism? It clearly is. If science discovers that there is nothing our color experiences have been tracking, then teleosemantics has to say that the mechanism that produces our color experiences is not a representational mechanism. That is, it is possible that color experiences are not representational states. (2013: 278)

The suggestion is that we could discover that our color experiences don't bear the relevant tracking relation to anything at all, and so, we could discover that we don't represent colors after all. (P1) would then be false, which would block the argument to color realism (which would also be false) in a way that is fully compatible with teleosemantics.

However, the same considerations that support the claim that color representations represent colors also support the weaker claim that they represent *something*. We can know from introspection that our representations of redness are not empty, that they in fact have contents. It is introspectively obvious that we think *something* when we think about colors. (This is something that Artiga and Millikan would deny. But surely the psychological role of color representations or other considerations that do not require realism about colors can help establish that color representations represent *something*.)

Incidentally, the suggestion that our (pseudo-)representations do not represent anything at all has unwanted consequences for the tracking theorist. While it allows her to deny color realism, it does so at the cost of making color anti-realism unthinkable if true. If there are no colors, and if our color "concepts" (or whatever we use to apparently think about colors) are supposed to get their content through tracking, then we would have no concept of color, and the thought *color realism is false* would not be thinkable.[22] We would not be able

[22]The only way to avoid this consequence would to be to claim that color (pseudo-)concepts are obtained not through tracking, but through composition of other representations that do

to represent to ourselves *what it is* that does not exist. Note that we cannot deny the existence of colors by justing thinking to ourselves that our "color" (pseudo-)representations fail to represent, since there is nothing that makes them pertain to colors, so, again, this thought will fail to tell us *what it is* that does not exist. One way to see this is to note that the thought that our (pseudo-)representations fail to represent is fully compatible with the existence of colors. Color anti-realism being unstatable if true is clearly absurd and a high price to pay to block the argument from (P1), (P2), and (REAL) to (C).[23,24]

Artiga's suggestion that teleosemantics can allow us to discover that we don't represent colors comes dangerously close to the claim that what we represent should be settled by theory, a claim that is problematic regardless of what we think about introspective self-knowledge. While some cases of mental representation might have to be settled by theory, many cases are such that we have some kind of theory-independent access to them. Tracking theories are theories of *mental representation*, not just theories about certain kinds of tracking re-

track something. But it does not seem that we represent colors from composition. In any case, we could run the same argument with some other (pseudo-)concept that is supposed to get its content directly from tracking.

[23] An anonymous reviewer has suggested that empty (pseudo-)representations might be cases of reliable misrepresentation, and so that taking color (pseudo-)representations to be empty would be a way for teleosemantics to allow for the reliable misrepresentation of color. However, empty (pseudo-)representations are not a kind of reliable misrepresentation, since reliable misrepresentation requires *representation* (by (RM1)), and empty (pseudo-)representations do not represent. Further, reliable misrepresentation requires *falsity* (by (RM2)), and empty (pseudo-)representations are arguably neither true nor false, since representing falsely requires representing.

[24] One might suggest that there is a difference between representing and seeming to represent (see Millikan 1984: 326-7). Perhaps all we can conclude from introspection and considerations of psychological role is that our color representations *seem* to represent colors. I'm not sure how to understand this claim other than as the claim that we represent that we represent colors. But then this suggestion faces the following dilemma: Either representing that we represent P requires representing P or it does not. (Representing that we represent P will require representing P on views of higher-order states on which lower-order states or their contents are embedded or otherwise involved in the higher-order states (see, e.g., Burge 1988), but such views of higher-order states are not mandatory.) If representing that we represent P requires representing P, then we can establish (P1) from the fact that we represent that we represent colors. If it does not, then this is presumably because the content *representing* P is not composed of other contents. But then we can run an amended the argument from (P1), (P2), and (REAL), to (C) where "redness" is replaced with "representing redness" and "colors" is replaced with "the representing of colors." Since this response accepts that we represent that we represent redness, it should accept that the new version of (P1) is true.

lations. This is why they are in competition with one another and with other theories of mental representation. Tracking theories aim at a certain target, mental representation, and attempt to account for it. While there are different ways of fixing reference on our target[25], we need to have some sort of theory-independent grip on it in order for the disagreement between different tracking theories, and different theories of mental representation more generally, to be a genuine disagreement.[26] If we had no theory-independent grip on mental representation, then teleosemantics and other kinds of tracking theories would not be in disagreement. They would each be theories of their favored kinds of tracking relations and nothing more. They might disagree on which tracking relations are most important, or which are useful for certain purposes, but they needn't be in competition with one another. All the tracking relations they specify could peacefully co-exist. Since tracking theorists seem to take their theories to be in competition, they should accept that they have a common target, which requires that there is a theory-independent way of fixing on this target.

With our theory-independent grip on mental representation come theory-independent ways of finding out what certain mental states represent. These ways might be based on introspection, intuition, or observations of inferences, brain states, or behaviors. Indeed, tracking theorists seem to accept that we have a theory-independent way of finding out about mental contents. This is clear in their discussions of the disjunction problem. The *disjunction problem* arises when a theory of mental representation incorrectly assigns disjunctive contents (e.g., *horse or skinny-cow-on-a-dark-night*) to mental representations that don't have disjunctive contents (e.g., HORSE) (see Fodor 1987: ch. 4). In order for the disjunction problem to actually be a problem, we need a theory-independent

[25] See Mendelovici MS: ch. 1 for discussion.

[26] In my view, our theory-independent grip on mental representation comes from introspection (see Mendelovici 2010: ch. 2, MS: ch. 1, Kriegel 2011: ch. 1). Other views are that it comes from folk psychology or cognitive science (see, e.g.,Fodor 1987).

way of knowing that the content of the relevant mental representations is not in fact disjunctive. Otherwise, we could just accept that certain theories claim that certain (or all) contents are disjunctive. That tracking theorists tend not to bite the bullet on the disjunction problem shows that they accept that there are ways of finding out what a mental representation represents independent of a theory of mental representation. Based on discussions of the disjunction problem, these ways seem to be largely based on intuition and introspection; it's supposed to just be *obvious* that HORSE doesn't represent *horse or skinny-cow-on-a-dark-night*.

All this is relevant to the argument for realism in two ways. First, this means that teleosemantics (or any theory of mental representation) is not free to dictate the contents of our mental states. We have theory-independent ways of finding out what we represent. While some cases of mental representation might have to be settled solely based on our theory, in many cases, pre-theoretical considerations constrain or completely inform us as to what is represented. In the case of perceptual experiences of colors and thoughts about colors, it is pre-theoretically clear that the relevant representations represent *something*. Suggesting that a tracking theory could inform us that we don't represent anything is no more convincing than suggesting it could inform us that HORSE represents *horse or skinny-cow-on-a-dark-night*. Tracking theorists should not bite the bullet on such cases, because biting these bullets is in tension with acknowledging that we have a theory-independent way of finding out what our mental representations represent, and denying that we have a theory-independent way of finding out what our mental representations represent is in tension with taking tracking theories of mental representation to actually be theories of mental representation.[27]

[27]It is not clear that the kind of self-knowledge provided by Millikan's theory (see fn. 18) provides us us the kind of theory-independent grip on mental representation required to adjudicate disagreements between different theories of mental representation on independent

Second, as long as the theory-independent ways of finding out what a representation represents don't require ascertaining that realism about a candidate represented property is true, tracking theories will license unacceptable arguments from premises like (P1) and (P2) to (C). In other words, even if we don't know (P1) through introspection, as I claim we do, we will still be able to bypass the standard considerations for ascertaining realism about a represented property as long as we are able to establish (P1) without checking the surfaces of objects. So, tracking theories license inappropriate consequences even if we deny introspective self-knowledge. Denying introspective self-knowledge is not enough to make moves licensed by (REAL) palatable; one must also deny other theory-independent ways of finding out what our mental representations represent. But denying that we have theory-independent ways of finding out what our mental representations represent is in tension with taking tracking theories

grounds. Millikan's theory of self-knowledge is a theory of how we come to know that two representations track the same thing, so it seems the insight it takes self-knowledge to provide ends up being insights onto what is tracked. If it turned out that her method delivered results that were at odds with her tracking theory, then presumably she would claim that her method delivered a mistake; after all, this method aims to find out whether we track the same thing on multiple occasions, so if its results come apart from what we in fact do track, we can conclude that it has made a mistake. This means that her method does not provide an independent means of validating the predictions of her theory; we could never use it find out that her theory was false.

Pietroski 1992 provides an imaginary case aimed at testing Millikan's teleosemantics on independent grounds: He asks us to imagine two species, kimus, and their only predators, snorfs. Kimus were originally color-blind, but by random mutation, one kimu has a representation R that is tokened in the presence of red light. In the morning, red light emanates from the top of a hill. Evolution eventually selected kimus that were fond of red light and hence would climb the hill every morning, thereby avoiding being eaten by snorfs, which happen to not be able to climb hills. Pietroski claims that Millikan's theory delivers the wrong result in this case: Her account predicts that R represents the lack of snorfs, a snorf-free zone, or something else to do with snorfs, but, he claims, whether R is a representation of red light, redness, something nice, or something else, one thing it certainly is not a representation of is anything to do with snorfs. In making this argument, Pietroski assumes that we have an independent way of knowing the contents of representational states (his favored way appeals to the role of mental representation in psychological explanations). Millikan bites the bullet on this objection, claiming that R does indeed represent something to do with snorfs. But if Pietroski's case does not count as evidence against teleosemantics from independent considerations pertaining to the content of representational states, it is not clear does.

See also Mendelovici and Bourget 2014, which argues that a naturalistic approach to mental representation requires more than reducing mental representation to the physical; it also requires being compatible with the theory-independent empirical evidence concerning what a representation represents.

to actually be theories of mental representation.

3 Conclusion

In this paper, I've overviewed, clarified, and improved various aspects of my argument from reliable misrepresentation against tracking theories. I've also presented and responded to certain objections made in Artiga 2013. If my arguments are sound, the argument from reliable misrepresentation escapes Artiga's objections, and tracking theories still need to allow for clean cases of mental representation.

I will close by summarizing my complaint against tracking theories, and against teleosemantics in particular, in an intuitive way: There could be cases in which we keep track of some worldly property, A (say, surface reflectance profiles), but we do this by representing "to ourselves" something else, B (say, primitive colors). Perhaps we need to keep track of A, but we do not need to know just what property it is, so it does not matter whether it is A or B that we represent. It might even be easier or more economical for us to represent B rather than A, perhaps because A is highly complex, while B is not. All this could be as it should be by any standard other than veridicality: the setup could be useful for us and our ancestors, and it could result is as strong a connection between our representations and B as we please. It need not be an accident or a byproduct of some other of our useful features. Teleosemantics, and tracking theories in general, inappropriately rule out this possibility on the basis of theory alone. But it is a live empirical possibility, one that should be left open by any theory, especially one claiming to be naturalistic.[28]

[28]Thanks to David Bourget, Marc Artiga, and two anonymous reviewers for helpful comments on previous drafts of this paper.

References

Artiga, M. (2013). Reliable misrepresentation and teleosemantics. *Disputatio*, (37):265–281.

Burge, T. (1988). Individualism and self-knowledge. *The Journal of Philosophy*, 85:649–663.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press, Cambridge, MA.

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

Kriegel, U. (2011). *The Sources of Intentionality*. Oxford University Press.

Martínez, M. (2013). Teleosemantics and Productivity. *Philosophical Psychology*, 26(1):47–68.

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

Mendelovici, A. (2013). Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies*, 165(2):421–443.

Mendelovici, A. (forthcoming-). Propositionalism without propositions, objectualism without objects. In Grzankowski, A. and Montague, M., editors, *Non-proposiitonal intentionality*. Oxford University Press, Oxford, UK.

Mendelovici, A. (MS). *Phenomenal intentionality: How to get intentionality from phenomenal consciousness*.

Mendelovici, A. and Bourget, D. (2014). Naturalizing intentionality: Tracking theories versus phenomenal intentionality theories. *Philosophy Compass*.

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

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

Millikan, R. G. (2000). *On Clear and Confused Ideas: An Essay About Substance Concepts*. Cambridge University Press.

Papineau, D. (1987). *Reality and Representation*. B. Blackwell.

Pietroski, P. M. (1992). Intentionality and teleological error. *Pacific Philosophical Quarterly*, 73:267–282.

Prinz, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. MIT Bradford, Cambridge, MA.

Rupert, R. D. (1999). Mental Representations and Millikan's Theory of Intentional Content: Does Biology Chase Causality? *Southern Journal of Philosophy*, 37(1):113–140.

Shea, N. (2002). Getting Clear About Equivocal Concepts. *Disputatio*, 13:34–47.

Shea, N. (2004). *On Millikan*. Wadsworth, Belmont.

Sterelny, K. (1989). Fodor's nativism. *Philosophical Studies*, 55(February):119–41.