CRÍTICA, Revista Hispanoamericana de Filosofía. Vol. 46, No. 136 (abril 2014): 69–86

PRINZ'S NATURALISTIC THEORY OF INTENTIONAL CONTENT

MARC ARTIGA
LOGOS/Universitat de Girona
marc.artiga@gmail.com

SUMMARY: This paper addresses Prinz's naturalistic theory of conceptual content, which he has defended in several works (Prinz 2000, 2002, 2006). More precisely, I present in detail and critically assess his account of referential content, which he distinguishes from nominal or cognitive content. The paper argues that Prinz's theory faces four important difficulties, which might have significant consequences for his overall empiricist project.

KEY WORDS: concepts, naturalism, empiricism, intentionality, counterfactuals

RESUMEN: Este artículo discute la teoría del contenido conceptual de Prinz, que él ha defendido en diversas obras (Prinz 2000, 2002, 2006). Más concretamente, presento en detalle y evalúo críticamente su teoría del contenido referencial, que él distingue del contenido cognitivo o nominal. El artículo argumenta que la teoría de Prinz tiene cuatro problemas importantes, que pueden tener consecuencias significativas para su proyecto empiricista.

PALABRAS CLAVE: conceptos, naturalismo, empiricismo, intencionalidad, contrafácticos

1. *Introduction*

Prinz is well-known for his outstanding contribution to the revitalization of concept empiricism in philosophy. Over the last decade, he has developed a sophisticated theory of concepts (e.g. Prinz 2000, 2002, 2006, 2008) which he has applied to other domains like a theory of emotions (2004) and a theory of morality (2007). In a nutshell, his version of conceptual empiricism is based on the idea that concepts are perceptually derived representations that he calls "proxytypes". Proxytypes are structured representations couched in modality specific formats that we employ in thought. As Prinz suggests, "all (human) concepts are copies or combinations of copies of perceptual representations" (Prinz 2002, p. 108).

In this paper I would like to present and discuss in some detail Prinz's naturalistic theory of conceptual content. Surprisingly, this is an aspect of his theory that has not been much discussed in the literature, even if it is a key premise in many of his arguments. For instance, when Prinz (2006) argues that we can perceive abstract entities (what he requires in order to explain the fact that we can think

about abstract entities), he supports his argument with a particular view of how conceptual content is determined. Similarly, he has also employed this account in his theory on emotions (Prinz 2004, pp. 93-94), among other places. Indeed, it is not unreasonable to think that the plausibility of his empiricist theory of concepts partially depends on whether perceptually derived representations can represent all the entities that we have concepts of. So his theory of representational content plays an essential role in the framework he wants to put forward. In this essay, I would like to show that his own theory of content determination falls prey to striking difficulties.

More precisely, here I will focus on Prinz's account of *referential* content, which Prinz distinguishes from something he calls "nominal content" (Prinz 2000) or "cognitive content" (Prinz 2002). A concept's referential content is the property, object or state in the world a concept refers to. For instance, the referential content of the concept DOG is *dog* (or, perhaps, *doghood*) and the referential content of the concept OBAMA is the individual *Obama*. There are three main reasons for focusing the discussion on referential content. First of all, Prinz provides an original theory of referential content, while he does not seem to offer any innovative account of nominal or cognitive content. Secondly, Prinz's theory of referential content is employed in many of his arguments in which a theory of content is playing an important role. Finally, an account of nominal content (which, in any case, Prinz has not developed in much detail; see Prinz 2000) will probably ride piggyback on a theory of referential content, so some of the problems of the former will probably carry over to any substantive theory of nominal content.

But, what is a *naturalistic* theory of referential content? The main goal of Prinz's theory of conceptual content is to explain in virtue of what process conceptual states acquire their (referential) content. That is, it seeks to explain why the concept DOG means *dog* rather than *cat* or *Paris*. Prinz wants to describe the process by means of which mental states come to have certain meanings. That the theory is naturalistic roughly means that the fact that a given state has a certain referential content has to be explained without appealing to other unanalysed intentional notions. In other words: the project is to explain in non-intentional terms (in terms of causation, information, covariance and the like) why certain states refer to certain entities. This is a traditional project in philosophy that has generated an extensive philosophical literature (for reviews, see Adams and Aizawa 2010 and Neander 2012). Here I would like to outline and discuss Prinz's contribution to this important topic.

2. *Prinz's Account*

First of all, it is worth pointing out that, in his empiricist approach to concepts, Prinz combines a non-atomistic theory of conceptual structure with an informational theory of content. That is, on the one hand, Prinz thinks that concepts are structured representations, composed of a set of perceptually derived representations. Concepts are individuated by taking into account a fuzzy network of associated representations. Nonetheless, at the same time, he holds that the referential content is determined by some sort of causal-informational connection that concepts have with their referents. Thus, while concepts are structured representations, their content is determined by a direct relation between representations and their representata.

More precisely, Prinz's account of content determination tries to combine Fodor's (1990) Asymmetric Dependence Theory and Dretske's (1981, 1986) Informational Theory, as he himself admits at several places (e.g. Prinz's 2006, p. 94). According to him, for a concept C to have X as its content (that is, for C to mean X) two conditions need to be met: (1) X must be C's *incipient cause* and (2) there has to be a *nomological covariance* between C and X. Let us look more carefully at each condition.

2.1. Incipient Causes

Prinz shares the widespread intuition that the naturalization of content should appeal to some kind of causal relation (see Stampe 1977). However, not any causal relation between an entity and a concept will do. For instance, a naïve causal theory that merely states that C means X iff C is caused by X would run into serious problems. First of all, this theory would entail that concepts have a highly indeterminate content. Certainly, snakes cause tokens of my concept SNAKE; but so do lizards or wooden sticks at dusk. In this case, the naïve causal theory would imply that my concept SNAKE means *snake or lizard or wooden stick*. The second striking difficulty of the naïve causal theory is that it fails to account for misrepresentation. Misrepresentation typically occurs when a concept is caused by something that is not in its extension. Since on this naïve theory any entity that causes a concept is immediately included in its extension, the most common situation that gives rise to misrepresentation is automatically ruled out.[1] Consequently, merely appealing to some

[1] Notice that the problem of error and the problem of indeterminacy are different. In principle, a theory can solve the former without solving the latter, if it allows for *some* cases of misrepresentation.

causal relation is insufficient. Prinz has to specify in more detail which is the causal relation that determines content.

Drawing on etiological theories of direct reference (Kripke 1980) and inspired by Dretske's (1981) appeal to a learning period, his suggestion is that the entity that causally originated the concept is specially important in determining reference. So he claims that the relevant cause must be the first one. This is why his first condition for content determination appeals to what he calls the "incipient cause": X is the incipient cause of the concept C iff X caused the formation of concept C. That is:

> INCIPIENT CAUSE: X is the incipient cause of C iff X is the first cause of C (i.e., X originated the creation of C).

According to Prinz, a necessary condition for C to mean X is that X has been the originating cause of the concept.

Still, the mere appeal to the incipient cause is insufficient for providing an adequate account of content (we will see that one of the main reasons has to do with problems of indeterminacy). For this reason, Prinz resorts to the tradition that postulates a covariance relation between a concept and its referent.

2.2. Nomological Covariation

The intuition that reference is determined by some sort of covariance is also common in the literature and has led to a range of different proposals (e.g. Dretske 1981, 1986; Rupert 2008; Eliasmith 2000). However, Prinz's notion of nomological covariance differs from other proposals in not being based on a covariance within the actual world, but across possible worlds. Prinz (2002, p. 241) defines nomological covariation in the following way:

> NOMOLOGICAL COVARIATION: Xs nomologically covary with concept C when Xs cause tokens of C in all proximate possible worlds where one possesses that concept.[2]

[2] Prinz sometimes adds a "ceteris paribus" condition, so that he sometimes defines nomological covariance in the following way: "Xs nomologically covary with concept C when, ceteris paribus, Xs cause tokens of C in all proximate possible worlds where one possesses that concept". I have removed this clause because any appeal to normal conditions or ceteris paribus conditions threatens to undermine the naturalistic credentials of the theory.

That is, John's concept DOG means dog partially because in all proximate possible worlds where John has DOG, tokens of this concept have been caused by dogs.

While NOMOLOGICAL COVARIATION connects with the tradition that seeks to naturalize content by appealing to a covariation between representations and their referents, Prinz's notion is different from other popular views. On the one hand, in contrast to the standard way of understanding "covariation" (on which, for instance, Dretske's or Rupert's account is based), NOMOLOGICAL COVARIATION is spelled out in counterfactual terms, and hence it is irrelevant how often X has correlated with C in the actual world. On the other hand, in contrast to other counterfactual theories such as Fodor's, NOMOLOGICAL COVARIATION does not take into consideration other possible causes of C. Whether C actually covaries with X only depends on the relation that holds between C and X in nearby possible worlds. Other possible or actual causes of C are not taken into account.

Now, it should also be clear that NOMOLOGICAL COVARIATION alone is too weak a relation for grounding semantic relations, because there are too many things mental states nomologically covary with. If in proximate worlds the transparent and colorless liquid that fills oceans and ponds is XYZ, then my concept WATER nomologically covaries with water (H_2O), but it also nomologically covaries with XYZ. More generally, anything that sufficiently resembles WATER in proximate worlds would be included in our concept WATER. If NOMOLOGICAL COVARIATION was the only condition for C to mean X, our concept WATER would mean *water or XYZ*. That would entail that concepts are highly disjunctive.

2.3. Incipient Theory

Consequently, Prinz (2002, p. 251) puts together these two notions (incipient cause and nomological covariation) in order to provide the necessary and sufficient conditions[3] for content determination:

INCIPIENT THEORY: X is the intentional (referential) content of C iff:

> 1. An X was the incipient cause of C, in accordance with INCIPIENT CAUSE.

[3] Let me mention that Prinz thinks these are necessary and sufficient conditions for the great majority of concepts. He wants to leave room for other concepts acquiring their content in a different way.

2. Xs nomologically covary with tokens of C, in accordance with NOMOLOGICAL COVARIATION.

There are two nice features in favor of this account. First of all, it seems to yield the right results in a wide range of cases. Take the concept TREE. On the one hand, we might reasonably suppose that we first developed this concept when we were confronted with a tree, rather than by seeing a cat or Obama. On the other, it seems that in all proximate worlds where I have this concept, trees still cause it. For instance, if we consider nearby worlds in which trees are a bit higher, or have a different color, or even worlds in which our visual system is slightly different, it seems that trees still cause my concept TREE. Thus, INCIPIENT THEORY gives the right result in many situations.

Secondly, this approach seems to be fully naturalistic. Only causal and counterfactual conditions are mentioned in INCIPIENT THEORY, so there is no intentional notion in the explanans. In that respect, it seems that Prinz's view should not raise any naturalistic qualms.

There are, however, some difficulties that seriously undermine the plausibility of this theory. Let me examine them carefully.

3. *Discussion*

I will present four objections to Prinz's view: the indeterminacy problem, the existence of ambiguous concepts, the phenomenon of meaning change and the question of circularity. I will also show that each of these objections is rooted in a central feature of Prinz's account, so that it is highly unlikely that any small modification of the account can provide a satisfactory solution.

3.1. Indeterminacy

It is well-known that many naturalistic theories suffer from indeterminacy problems (Fodor 1990; for some replies, see Agar 1993; Price 1998, 2001; Millikan 2004). Knowingly, Prinz provides an original reply to the traditional indeterminacy problem, that seems to successfully refute this objection. In this section, I would like to show that Prinz's original solution gives raise to an indeterminacy problem at another locus. The upshot is that the Incipient Theory is not immune to some version of the indeterminacy problem and probably lacks the resources for dealing with it. Let us go step by step.

A general way of stating the problem of indeterminacy is the following: a theory suffers from the indeterminacy problem if the theory entails that there are many entities represented by a given state, while common sense and science assume that it has a much more determinate content. Think, for instance, about John's MONARCH concept, that is, the concept that we would naturally attribute to John, which seems to unambiguously refer to monarch butterflies (John uses it when he sees a monarch, he calls it "Monarch", he associates with it the property of being a flying animal, and so on). Following Prinz, we can reasonably assume that the incipient cause of John's concept was a monarch and that this concept nomologically covaries with monarchs. However, monarchs are butterflies (indeed, this is a good candidate for being a necessary truth). So if a monarch was the incipient cause of John's concept, so was a butterfly. Thus, if condition 1 is satisfied by a monarch it is also satisfied by a butterfly. Similarly, if in all proximate possible worlds monarchs cause tokens of John's concept, butterflies also do (again, because monarchs are butterflies). So condition 2 is also satisfied by butterflies. Therefore, John's concept MONARCH means *monarch or butterfly*.

Similar results can be obtained with a wide range of properties: insect, animal, . . . The consequence seems to be that INCIPIENT THEORY entails that the content of John's concept is *monarch or butterfly or insect or*. . . In fact, even the property of being a monarch-looking thing causes troubles, since condition 1 and 2 of INCIPIENT THEORY seem to be satisfied: if a monarch was the incipient cause of John's concept, a monarch-looking thing probably was as well, and if monarchs cause John's mental state in the actual world, a monarch-looking thing will probably cause John's mental state in close possible worlds. The consequence of this analysis is that John's concept has a highly indeterminate content, which starkly contrasts with the original assumption that John had the concept that referred to monarchs (and only monarchs).[4]

Notice that a similar problem concerns any concept, so the objection generalizes: for any concept C, INCIPIENT THEORY entails that C has a highly indeterminate content. Consequently, even if appealing to incipient causes enables the theory to avoid including entities existing in proximate worlds that resemble very much the entities in

[4] Of course, one could say that, in this case, John's concept is not the concept MONARCH, but the concept MONARCH OR BUTTERFLY, etc. . . . In that case, the objection should be cashed out in the following terms: Prinz's theory entails that John lacks the concept MONARCH, as well as a huge amount of other concepts: TREE, WATER, GOLD, etcetera.

the actual world (such as H_2O and XYZ), there are still many sources of indeterminacy that INCIPIENT THEORY cannot exclude.

Interestingly, Prinz sometimes seems to be suggesting that, as stated, INCIPIENT THEORY can already deal with the serious problems of indeterminacy (Prinz 2002, p. 241). Nonetheless, at other places he adds further conditions in order to deal with this objection.[5] In particular, in Prinz (2002, pp. 242–243) he tries to solve what he calls the "semantic-marker" problem, which basically is a version of the indeterminacy problem suggested earlier. He claims that three further conditions need to be added to INCIPIENT THEORY in order to determine whether a concept refers to a natural kind, an individual or an appearance property (such as *being a monarch-looking thing*). These conditions are labeled "Semantic Markers":

SEMANTIC MARKERS

> (a) C is a kind concept if, had Xs looked different than they do, they would still cause tokens of C.
>
> (b) C is an appearance concept if, had Xs always looked different than they do, they would not cause tokens of C.
>
> (c) C is an individual concept if, were the subject presented with objects that appear exactly like X, at most one of those objects would cause tokens of C.

If we focus on a) and b), the idea is the following: consider the set of proximate worlds where Xs look different than they look in the actual world. If in these worlds Xs still cause C, then C is a kind concept. If they do not, then C is a concept of an appearance (a concept of X-looking thing). c) tries to apply the same idea to the case of individuals. Prinz's suggestion is that if SEMANTIC MARKERS is added to the two conditions of INCIPIENT THEORY, we would get an account that attributes determinate contents to concepts.

The first important thing to notice about this proposal is that, in contrast to what Prinz claims, a) b) and c) are in fact new conditions that should be added to INCIPIENT THEORY, rather than embellishments of condition 2. There is an easy way to see why this is so: whereas condition 2 of INCIPIENT THEORY states that we should consider proximate worlds, clauses a), b) and c) appeal to

[5] He presents these additional conditions as slight refinements of condition 2 in INCIPIENT THEORY. However, I will argue that they constitute new requirements.

those worlds where things look a different way, which might be very distant worlds. For instance, if in all proximate worlds Xs still look the same way, in order to assess whether a), b) or c) hold we might have to take into account distant worlds. In contrast, in order to see whether condition 2 holds, we should only consider proximate worlds. This shows that this solution to the semantic markers problem brings a new set of clauses into the definition.

Secondly, there is an obvious problem with simply adding SEMANTIC MARKERS to the previous definition: even if this proposal succeeded, it would only provide a recipe for distinguishing concepts of kinds, concepts of appearances and concepts of individuals, while the problem of indeterminacy is much more widespread. Monarchs, butterflies and insects are natural kinds and they all generate the indeterminacy problem, so merely adding the counterfactual conditions stated in SEMANTIC MARKERS to INCIPIENT THEORY will not tease apart concepts of monarchs, butterflies and insects. Semantic markers are not fine-grained enough for the task at hand. Therefore, Prinz's theory seems to fall prey to the indeterminacy problem, even if semantic markers are added.

Now, whereas I think Prinz's appeal to three semantic markers fails to solve the indeterminacy problem, I would like to explore a possible reply on behalf of Prinz's approach. Basically, the idea is to generalize the strategy of semantic markers suggested in a), b) and c) in order to rule out any inadequate properties. The proposal is the following: for any properties X and Y that satisfy conditions 1 and 2 of INCIPIENT THEORY (that is, for any two properties that cause problems of indeterminacy), consider the most proximate worlds in which one is instantiated but not the other (say, X is instantiated, but not Y). If in those worlds, X still causes tokens of the concept, then C means X (and not Y). If it does not, then C does not mean X. That is:

BETTER SEMANTIC MARKERS

> For any properties X and Y that satisfy 1 and 2 of INCIPIENT THEORY, consider the set of proximate worlds where one has the concept C and Xs are not Y:
>
> 1. If Xs still cause tokens of C, C represents X (and not Y).
>
> 2. If Xs do not cause C, C does not represent X.

That is, in order to know whether John's concept refers to monarchs or butterflies, BETTER SEMANTIC MARKERS tells us to consider the possible worlds where there are butterflies but no monarchs; if in those worlds butterflies still cause tokens of John's concept C, then it is a concept of butterfly (and not of monarch); if butterflies do not cause C, then John's concept is not a concept of butterfly.[6] Similarly, in order to know whether C is about monarchs or monarch-looking things, look at the most proximal worlds where monarchs are not monarch-looking things;[7] if monarchs still cause C, then C is about monarchs. Otherwise, C is about monarch-looking things.

Now, if we add BETTER SEMANTIC MARKERS to the original theory formulated in INCIPIENT THEORY, we get the following account:

BETTER INCIPIENT THEORY

X is the intentional (referential) content of C iff:

1. An X was the incipient cause of C, in accordance with INCIPIENT CAUSE.
2. Xs nomologically covary with C, in accordance with NOMOLOGICAL COVARIATION.
3. For any properties X and Y that satisfy 1 and 2, consider the set of proximate worlds where one has the concept C and Xs are not Y:

   (a) If Xs still cause tokens of C, C represents X (and not Y).

   (b) If Xs do not cause C, C does not represent X. (BETTER SEMANTIC MARKERS)

I think this is a better proposal than the previous one, and BETTER SEMANTIC MARKERS provides a finer-grained reply to the problem of indeterminacy than Prinz's distinction between kind, appearance and individual concepts. Unfortunately, I think that even this refined version of INCIPIENT THEORY utterly fails to solve the indeterminacy problem. There is an important difficulty that this solution to the semantic marker problem cannot deal with.

[6] Of course, in order to do that, concepts should be individuated narrowly.

[7] "Being monarch-looking" refers to the property of looking the way monarchs look in the actual world. If the property referred to the different ways monarchs look in different worlds, there would be no world at which monarchs do not instantiate the property "being monarch-looking".

First of all, remember why we abandoned the naïve causal account sketched earlier (which claims that a state represents whatever causes it): since in the actual world many different entities cause mental states, that would yield a highly indeterminate content. Since my concept SNAKE is caused by snakes, lizards and even wooden sticks at dusk, all these entities would figure in the content of the concept.

Now, the problem of the kind of reply suggested by Prinz (either SEMANTIC MARKERS or BETTER SEMANTIC MARKERS) is that it assumes that by merely moving to other possible worlds, we will be able to distinguish the right cause from the wrong causes; however, this is far from clear. We saw that *in the actual world* some of the things that cause my concept MONARCH are not monarchs, so why should we think proximate possible worlds are any different in that respect? It seems that, prima facie, if we move to nearby possible worlds where a subject has the concept C the same situation will probably arise. In those possible worlds where I have the concept MONARCH (narrowly individuated) many things that are not monarchs also cause tokenings of my concept. So merely moving to other possible worlds does not enable us to distinguish the right from the wrong causes.

Hence, the problem is the following: even if John's concept means *monarch*, in some of the possible worlds where butterflies are not monarchs, butterflies cause tokens of John's concept, so condition 3 will not rule out *butterfly* from the content. More generally, if we move to those possible worlds where properties X and Y are not coinstantiated, we will probably find out that in some of these worlds X (but not Y) cause C and in some other worlds Y (but not X) cause C. Concepts are also caused by the wrong entities in other possible worlds. So, BETTER SEMANTIC MARKERS will not help us in determining content for the same reason that the naïve causal account did not work: the fact that misrepresentation is possible shows that content cannot be determined by what causes a certain mental state. And this claim holds here and in other possible worlds.[8]

The reason Prinz's theory faces a misrepresentation problem at that particular point and not earlier is that if we focus on the ac-

[8] This is also the reason why similar proposals that rely on causal relations holding in close possible worlds will probably fail. For instance, merely appealing to the entity that *generally* causes tokens of a concept in close possible worlds, or the entity that *mostly* causes tokens of a concept in close possible worlds will not solve the problem. My MONARCH concept can be very often and systematically caused by entities that are not monarchs in close possible worlds: butterflies, viceroys, toys and so on.

tual world, he has an satisfactory reply: only the first cause (the incipient cause) determines content. So (if we assume that concepts are always firstly caused by instances of their referents), he has a way of distinguishing misrepresentations from true representations. In contrast, when he appeals to causal relations holding in other possible worlds in order to determine the content of the concept in the actual world, his theory yields the wrong results. In other possible worlds, anything can cause my mental state.[9] Consequently, even if MONARCH means *monarch* and not *butterfly*, if we move to worlds where monarchs are not butterflies, we will probably find that some butterflies (which are not monarchs) still cause tokens of the concept and in some of these worlds they do not. As a result, the fact that X and not Y causes tokens of a concept C at those worlds where X and Y are not coinstantiated cannot help to determine content.[10] And, again, notice that this problem generalizes: any entity that can

[9] Furthermore, in that case, he cannot plausibly modify BETTER SEMANTIC MARKERS so that only the incipient cause in other possible worlds is relevant. That would surely be too strong a condition; even if we grant that in the actual world my concept TREE was first caused by a tree, it seems that there are many possible worlds were trees are not the incipient cause of my concept TREE.

[10] In a previous version of the theory, he offers a slightly different condition. In Prinz (2000, p. 13) he claims that X nomologically covaries with Y iff (1) Xs cause Ys in all proximate nomologically possible worlds, and (2) when they do so, they do so in virtue of being Xs. At first glance, one might think this version of the theory avoids the problem of indeterminacy I am pointing out, since it provides a way of distinguishing the right causes from the wrong causes in other possible worlds: Prinz can simply appeal to the fact that, in any world, tokens of MONARCH are caused by an entity *in virtue of* its being a monarch and not *in virtue of* its being a butterfly.

However, this proposal raises naturalistic qualms. The problem lies in the fact that the relation *in virtue of* is doing all the work and should be specified further. More precisely, one might worry that we presume that X causes C in virtue of its being X (i.e. that condition 2 is true) because we are presupposing precisely what we are trying to explain, namely that Y means X (this issue will be extended below, in the section "circularity").

Think about it in the following way: if one were allowed to appeal to X causing Y in virtue of being an X, then nothing like incipient causes or nomological covariance would be required. One could just claim that Y means X iff Xs cause tokens of Y in virtue of being X. And the obvious problem with this view is that it is completely uninformative. In this approach, the notion "in virtue of" seems to merely label the relation we are trying to explain, rather than offering an explanation.

Indeed, Prinz himself (2000, p. 13) admits that this notion should be explained, and claims that "Xs cause Ys in virtue of being Xs when (1) when an a that is X causes Y, if a were not X, it would not cause Y or (2) when an a that is X causes Y, there is no other nomologically sufficient cause of Y". However, this way of cashing out the relation *in virtue of* shows that this version of the theory also suffers from the indeterminacy problem. Even if my concept MONARCH means *monarch*, some

be misidentified in the actual world can also be misidentified in other possible worlds.

Therefore, I think Prinz's theory cannot solve the indeterminacy problem.

3.2. Ambiguity

Secondly, not only the nomological condition, but also the incipient cause condition runs into problems.

There is a set of counterexamples that Prinz has not appropriately addressed: ambiguous concepts. A consequence of INCIPIENT THEORY seems to be that (non-deferential) concepts can never have ambiguous contents. Suppose I have a concept C that I equally apply to trees of kind A and trees of kind B, and suppose I have never heard about these trees, nor do I intend to defer the fixation of meaning to experts (so, suppose this concept C is not deferential). In that case, my concept would either mean *tree of kind A or tree of kind B*, depending on the entity that first caused it. That is an implausible result, for at least two reasons. On the one hand, it seems that, if I have always consistently and repeatedly applied a concept C to two entities, the intuitive result should be that this concept is ambiguous. It should mean something like *being tree A* or *being tree B*. More generally, it is plausible that in fact some of our concepts are ambiguous, and language does not seem to be required for having them (Millikan, 2000). Secondly, whether a tree of kind A or a tree of kind B was the first cause seems to be a matter of luck, but the content of my concept C does not seem to depend on such a chancy event. In this case the strictness OF INCIPIENT CAUSE makes it difficult to account for these cases where meaning is disjunctive.[11]

One could reply that we are unduly restricting the set of causes. In particular, one might suggest that we should also consider disjunctive causes, such as *being tree A or tree B*. Certainly, it seems that, in the example just given, *being tree A or tree B* is the incipient cause of the concept and this relation is robust, so this suggestion seems to get the example right. However, why should we consider a disjunctive cause rather than the simplest one? After all, *being tree*

butterflies cause tokens of this concept in the actual world and in other possible worlds, so both monarchs and butterflies satisfy (1) and (2).

[11] I focus on non-deferential concepts because Prinz (2002 pp. 254–255) has provided an interesting reply to this problem for deferential concepts: since in the case of deferential concepts there is a community involved, there might be different incipient causes for the same concept (Prinz 2002, pp. 254–255). This creative suggestion cannot be used in the case of non-deferential concepts.

A and *being tree B* are natural kinds (or, at least, they are more natural properties than *being tree A or being tree B*). It is hard to think of any principled reason for preferring the disjunctive property that cannot be accused of being ad hoc.

Even more troubling, the main problem of resorting to disjunctive causes in a naturalistic project is that they quickly multiply out of control. For instance, once disjunctive properties such as *being tree A or being tree B* are accepted, what prevents us from seriously considering the property *being H_2O or being XYZ* (which is the incipient cause and nomologically covaries with our concept WATER)? And why not considering bizarre properties such as *being H_2O or being a unicorn*? Thus, taking this option would solve a problem at the cost of creating a more difficult one. Merely appealing to incipient causes does not allow for any kind of disjunctive content, but accepting disjunctive causes makes all concepts highly disjunctive. Both results are clearly unsatisfactory.

3.3. Meaning Change

The third difficulty concerns change of meaning. Prinz considers a case where an alligator causes the creation of John's concept, but (due to an accident) afterward John consistently deploys this concept for crocodiles (suppose John happens to never see an alligator again). We can even imagine that, after many years of gathering information, at the end of his life John becomes a crocodile expert; he knows lots of things about them, can identify them very quickly, and so on. INCIPIENT THEORY seems to be committed to the view that the concept has always been wrongly applied to crocodiles. Since it was first caused by an alligator, it can only mean *alligator* (if anything). This is an extremely counterintuitive result. Perhaps my concept BUTTERFLY was first created by seeing a butterfly puppet, but since I have been applying this concept to butterflies for the rest of my life, it is not unreasonable to think it refers to butterflies and not to puppets.

Prinz tries out two kinds of replies. One is to say that at some point I created a new concept CROCODILE, whose incipient cause was indeed a crocodile. But why should we think John creates a new concept? We can stipulate there is no intention of creating a new concept; it just happens to be the case that John's concept becomes overwhelmingly correlated with crocodiles. Why should we think that at some (arbitrary) point John's concept became a different one? Notice that, since INCIPIENT THEORY requires an incipient cause, Prinz

has to assume that there is a particular point (say, after seeing 56 crocodiles) at which a new concept is created. So he is committed to accept that the 55th token of the concept was a wrong application of ALLIGATOR to a crocodile, and the next token was already a right application of the concept CROCODILE to a crocodile. That looks very implausible.

A second strategy Prinz pursues in order to account for meaning change is to bite the bullet and accept that, even if a subject has applied a concept C all his life to an entity X, if it was first caused by Y, the concept refers to Y. Nonetheless, in an attempt to make this result more palatable, Prinz compares concepts to artifacts; if a subject creates a tool for swattering flies, it will always be a fly swatter, no matter whether it is used as a paperweight or as a can opener (Prinz 2002, p. 254). Origins matter, he claims. Similarly, if John creates his concept when he was perceiving an alligator, then it is a concept of alligator.

Now, I think there are two crucial points at which this comparison breaks down. First, arguably, a tool is a fly swatter because someone *intended* it to be a fly swatter. That explains why something can be a fly swatter, even if it has never been used to swat any fly. The function of artifacts is (at least, partially) derived from the intentions of their designers. However, that feature starkly contrasts with the naturalistic account of concepts we are trying to provide. If the semantic properties of concepts derived from the intentional properties of the subject, nothing like nomological covariance or incipient causes would be relevant. That would undermine BETTER INCIPIENT THEORY. Furthermore, the semantic properties of concepts would be primarily explained by appealing to further intentional properties, so Prinz would be offering no *naturalistic* theory of intentional states.

Secondly, the comparison with artifacts is also inadequate because artifacts can have many functions at the same time (they can *be* many things), so we have no problem in saying that the fly swatter is still a fly swatter (even if it has never been used as such), because we can also say that, in addition, it is a paperweight. However, this is not the case with concepts: BETTER INCIPIENT THEORY entails that my concept ALLIGATOR is about alligators and not about crocodiles.

Consequently, Prinz's account has also counterintuitive results with respect to concepts that were created in a certain way but change their meaning.

3.4. Circularity

Finally, I would like to raise a general worry concerning this sort of approach. The last problem of BETTER INCIPIENT THEORY (and INCIPIENT THEORY) I would like to consider is that we lack a non-intentional justification of why condition 2 (which appeals to NOMOLOGICAL COVARIANCE) should hold. Let me elaborate on that point.

Consider first the relation between trees and the concept TREE. Why do trees in most proximate worlds systematically cause tokens of the concept TREE? Well, a plausible explanation is that this is true precisely because TREE means *tree*. The worry, of course, is that if the counterfactual claim is true in virtue of TREE meaning *tree*, then one is not allowed to appeal to this counterfactual condition in offering a naturalistic account of referential content. Circularity threatens.

Let me put the point in a different way. The truth of counterfactual statements is usually thought to be grounded in properties and relations holding in the actual world (at least, that seems to be a usual assumption of naturalistic accounts). For instance, consider the following counterfactual: *If Obama had not won the elections, Romney would have become the U.S. President.* Unless one is a modal realist (Lewis 1986), a naturalist will probably think that this counterfactual statement is true because of certain properties and causal relations holding in the actual world, probably involving Obama, Romney, certain social facts and so on. Similarly, some counterfactuals are true in virtue of certain intentional facts that hold in the actual world. As a consequence, if one is trying to naturalize an intentional relation by appealing to counterfactuals, one should be careful not to rely on counterfactual statements whose truth depends on the very same intentional relations one is trying to naturalize. This, I think, should be obvious.

Now, the trouble with naturalistic theories of content that appeal to counterfactuals such as Prinz's is that they might be assuming these intentional facts in the explananda. The suspicion is that counterfactual statements that are supposed to play a role in the naturalization of intentional content might be true in virtue of certain intentional relations holding in the actual world —namely those intentional relations that they are seeking to naturalize. So, unless Prinz specifies which properties and relations in the actual world ground the truth of these counterfactuals, the naturalistic credentials of this account will be dubious. Since no such characterization is provided,

one might reasonably suspect that this account might be relying on the intentional facts that it is trying to explain. Certainly, the concept TREE is caused by trees in close possible worlds; but this is true in virtue of the fact that TREE means *tree* in the actual world.

Indeed, I think this objection can probably be extended to all naturalistic accounts that rely on counterfactuals. In particular, it could probably be argued that Fodor's (1990) Asymmetric Dependence Theory suffers from the same problem. That should come as no surprise; given that Prinz admits that his theory is based on Fodor's view, one should expect to find some of its virtues as well as some of its flaws.

4. *Conclusion*

In this paper I have argued that Prinz's naturalistic theory of conceptual content faces four daunting difficulties. The objections put forward suggest that one aspect of his original and groundbreaking theory of concepts should be seriously reconsidered.

Even more importantly, given the relevance of this naturalistic theory of content in his overall naturalistic project (including the study of emotions and morals), abandoning this account of referential content will probably have major consequences for his views in other fields. If the arguments of this paper are sound, they open the door to some important revisions of Prinz's empiricist project.[12]

REFERENCES

Adams, F. and K. Aizawa, 2010, "Causal Theories of Mental Content", in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, Stanford University; available at: <http://plato.stanford.edu/archives/spr2010/entries/content-causal/> [01/09/2013].

Agar, N., 1993, "What Do Frogs Really Believe?", *Australasian Journal of Philosophy*, vol. 71, nos. 1–2, pp. 1–12.

Dretske, F., 1986, "Misrepresentation", in R. Bogdan (ed.), *Belief, Form, Content and Function*, Oxford University Press, New York, 1986, pp. 17–36.

[12] I would like to thank Jesse Prinz, David Pineda, and the audiences of the XXII SIUCC Workshop in San Sebastian and the LOGOS Graduate Reading Group for helpful comments. This work was supported by the scholarship BES–2008–005255 from the Spanish Ministry of Science and Innovation (MICINN), the Research Projects "The Naturalization of Subjectivity" (ref. FFI2010–15717), "Modal Aspects of Materialist Realism" (ref. HUM2007–61108) and Consolider-Ingenio project CSD2009–00056.

Dretske, F., 1981, *Knowledge and the Flow of Information*, MIT Press, Cambridge, Mass.
Eliasmith, C., 2000, "How Neurons Mean: A Neurocomputational Theory of Representational Content", unpublished Dissertation, Washington University in St. Louis.
Fodor, J., 1990, *A Theory of Content and Other Essays*, MIT Press, Cambridge, Mass.
Kripke, S., 1980, *Naming and Necessity*, Blackwell, Oxford.
Lewis, D., 1986, *On the Plurality of Worlds*, Blackwell, Oxford.
Millikan, R.G., 2004, *Varieties of Meaning*, MIT Press, Cambridge, Mass./London.
———, 2000, *On Clear and Confused Ideas*, Cambridge University Press, Cambridge.
Neander, K., 2012, "Teleological Theories of Mental Content", in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, Stanford University; available at: <http://plato.stanford.edu/archives/spr2012/entries/content-teleological/> [01/09/2013].
Price, C., 2001, *Functions in Mind*, Oxford University Press, Oxford.
———, 1998, "Determinate Functions", *Noûs*, vol. 32, no. 1, pp. 54–75.
Prinz, J., 2008, "Regaining Composure: A Defense of Prototype Compositionality", in W. Hinzen, M. Werning, and E. Machery (eds.), *The Oxford Handbook of Compositionality*, Oxford University Press, Oxford.
———, 2007, *The Emotional Construction of Morals*, Oxford University Press, Oxford.
———, 2006, "Beyond Appearances: The Content of Sensation and Perception", in T.S. Gendler and J. Hawthorne (eds.), *Perceptual Experience*, Oxford University Press, Oxford, 2006.
———, 2004, *Gut Reactions: A Perceptual Theory of Emotion*, Oxford University Press, Oxford.
———, 2002, *Furnishing the Mind: Concepts and Their Perceptual Basis*, MIT Press, Cambridge, Mass.
———, 2000, "The Duality of Content", *Philosophical Studies*, vol. 100, no. 1, pp. 1–34.
Rupert, R.D., 2008, "Causal Theories of Mental Content", *Philosophy Compass*, vol. 3, pp. 353–380.
Stampe, D., 1977, "Towards a Causal Theory of Linguistic Representation", *Midwest Studies in Philosophy*, vol. 2, pp. 42–63.

Received: August 27, 2013; revised: November 24, 2013; accepted: December 10, 2013.