Information and Metacognition

M.A. Sebastián, M. Artiga[*]

Abstract

Representations are daily postulated in mainstream neuroscience. Nonetheless, few have actually tried to offer a general theory of representation based on those practices. Recently, some approaches have attempted to develop this idea and defend that the relation of representation can be explained in purely informational terms. In this paper we argue that such informational theories cannot provide a satisfactory account of the relation of representation. In particular, we will show that they cannot accommodate the existence of metarepresentations, which play a central role in the explanation of certain cognitive abilities.

1 Introduction. Representation

Cognitive science attempts to explain how our cognitive system works. Although over the years there have been different ways of approaching this question, the mainstream view in still maintains that our mind is a representational system. Adherents of this view claim that the best way to explain cognition is to posit the construction of internal representations. Thus, to understand current practice in cognitive science, we need to get a better grasp of the nature of these entities: we need a theory of representational content.

In philosophy, this has traditionally been known as the problem of intentionality. And, although cognitive scientists in general, and neuroscientists in particular, do not usually address this problem directly, they seem to implicitly assume a set of intuitive conditions that are sufficient—or even necessary—for a state to qualify as a representation and to possess a determined representational content. In this paper we would like to examine recent attempts to turn this intuitive methodology into a full-blown naturalistic theory of representation. As we will see, these approaches heavily rely on the idea that information, understood as some form of statistical dependence, is the clue to understand representations. Of course, the idea that we can explain representations by appealing to some sort of information is not new, and can be traced back at least to Dretske (1981). Nonetheless, recent approaches are appealing for at least two reasons. First, they seem to solve the main difficulties faced by Dretske's informational theory. Secondly, and even more interestingly, they seem to capture the intuitive criteria employed by neuroscientist when they claim, for instance, that certain neuronal activation in a determinate cortical area represents a particular stimulus. Since they achieve these goals by modifying Dretske's original proposal in different ways, we will call this family of approaches 'Scientifically Guided Informational Theories' (SGITs). Some form or other of SGIT has been defended by Usher (2001), Eliasmith (2000, 2003), Rupert (1999), Skyrms (2010) or Scarantino (2015).

[*] This work is fully collaborative. Authors appear in random order.

Interesting as SGITs are, in this paper we argue that this kind of theories lack the resources to make a fundamental distinction that is at the core of many cognitive theories: the difference between those representations that have other representations as their object —i.e. metarepresentations— and those representations that are merely caused by other representations but have external stimuli as their object. Since representations of external stimuli and metarepresentations involve the same kind of relation—namely that of representation—, but play different and indispensable roles in our cognitive architecture, a satisfactory theory of representation needs to make room for such a distinction. If we are right, though, SGITs are unable to make it.

The paper is organized as follows: section 2 presents SGITs and section 3 clarifies the relevance of metarepresentations in our cognitive architecture. In section 4, we develop the idea that SGITs are unable to account for the difference between metarepresentations and representations of external stimuli and we consider some objections. Our argument is supposed to show that content cannot be fully determined solely in terms of statistical dependance relations. In section 5, we briefly discuss whether the notion of teleological function could help these approaches to solve this problem.

2 Informational Theories

Scientifically Guided Informational Theories (SGITs) are naturalistic theories of content. The main goal of these theories is to show how representations fit our scientific worldview. More precisely, they try to explain what it is for a state to be a representation and how its content is determined by appealing to non-representational states and processes. If that project could be carried out successfully, it would provide a solution the classical problem of intentionality. The nature of representational phenomena would be finally understood.

The particular answer SGITs gives to this challenge connects with the long-standing informational tradition. The key feature of Informational theories of content is that they seek to account for representations by resorting to some sort of informational relation.[1] One of the first and better known informational theories of content was Dretske's (1981), who tried to analyze semantic content by appealing to informational content and defined informational content in terms of probability relations. More precisely, according to his approach a state R carries information about another state S iff given certain background conditions $P(S \mid R) = 1$. While the idea of explaining semantic properties in terms of information was revolutionary and very influential, there were two deep problems with Dretske's proposal. First of all, in the natural world it is extremely difficult to find two different states such that the existence of one of them makes the other state certain (even if certain background conditions are assumed). This consequence made the theory unrealistic. Secondly, this approach was incompatible with one of the defining characteristics of representational states, namely that they can sometimes misrepresent. On Dretske's approach, a state represents another one only if both obtain, so a typical case of misrepresentation (which usually involves an existing state representing a non-existing one) is rendered impossible.[2] These and other problems lead most people to think that a satisfactory

[1] Although some of these theories have not explicitly been formulated in terms of 'information', we classify them under the label 'informational theories' because all of them try to accommodate representational relations by appealing to probability relations. At least in a certain way of understanding this notion, a certain amount of correlation is sufficient for an entity to carry information about another entity (Floridi 2010).

[2] Dretske tried to solve these problems by distinguishing a learning period (in which misrepresentation is still im-

informational theory of content was unworkable.

However, this situation has recently changed and some new informational theories are being put forward by philosophers and psychologists. Of course, these approaches are aware of the problems faced by previous theories in the same tradition, and for this reason define and use the notion of information in a slightly different ways. The main modification is the rejection of the requirement that $P(S \mid R) = 1$, which was the key assumption that caused the theory to be too unrealistic and make misrepresentation impossible. However, dropping this assumption is not without costs. In particular, which probability should then be required for a state to represent another state? Any lower standard would seem arbitrary. Furthermore, given the wide range of different representations one can find in the natural world, any arbitrary criterion will probably leave some real representations out and let some non-representational states in. To address these concerns, the strategy pursued by new informational approaches is to appeal to relative probabilities. Accordingly, what is relevant is not how much the representation raises the probability of another state, but whether it raises more the probability of a certain state than the probability of others. This is the central idea that has been developed in various ways by different authors.

Since a joint consideration of all SGIT would be extremely complex, for the sake of simplicity we will focus on a particular approach. Nonetheless, after presenting our objections, we will show how these problems probably extend to other SGITs (see section 4.4). More precisely, here we will concentrate on Usher (2001), because he defends an informational theory based on statistical dependence relations, provides a particularly clear approach and is explicitly motivated by research in cognitive science. Furthermore, his view seems to capture the intuitions expressed by Eliasmith (2000, 2005b, 2005a) and Rupert (1999), among others.

Usher (2001) claims that his account is based on Shannon's (1948) notion of *mutual information*. The core idea behind this concept is that a signal X provides information about some random variable Y just in case the presence of the X reduces the uncertainty of Y. In other words, just in case $P(Y \mid X) > P(Y)$. Shannon provided a precise mathematical definition of mutual information between two sets, that can be easily extended to calculate the mutual information between two states. In particular, the mutual information between X and Y (expressed as 'MI (X;Y)') is defined by the following formula:

$$MI(X;Y) = \log_2\left(\frac{P(X \cap Y)}{P(X)\,P(Y)}\right).$$

Therefore, the mutual information depends on the ratio $\frac{P(X \cap Y)}{P(X)P(Y)}$, which is identical to $\frac{P(X|Y)}{P(X)}$ and to $\frac{P(Y|X)}{P(Y)}$ (by Bayes' rule). Recall, however, that one of the central motivations of SGITs is that representational content cannot just be determined by the fact that a the mutual information between two variables reaches a certain threshold (this is the key point of departure from classical informational theories). Following this line of reasoning, Usher's proposal is that R represents S iff (1) the mutual information R carries about S is greater that the information it carries about any other entity and (2) the mutual information between S and R is greater than the information S carries about any other representation. More precisely:

1. $MI(R_i; S_i) = \frac{P(R_i|S_i)}{P(R_i)} > \frac{P(R_i|S_j)}{P(R_i)} = MI(R_i; S_j)$, for all $j \neq i$

possible) from a post-learning period, but it is generally agreed that this proposal probably cannot solve any of these difficulties.

2. $MI(R_i; S_i) = \frac{P(S_i|R_i)}{P(S_i)} > \frac{P(S_i|R_j)}{P(S_i)} = MI(R_j; S_i)$, for all $j \neq i$

Because of the identical denominator, these expressions can be simplified in order to provide a more concise definition of Usher´s informational theory:

Info R_i represents S_i iff for all $j \neq i$

1. $P(R_i \mid S_i) > P(R_i \mid S_j)$
2. $P(S_i \mid R_i) > P(S_i \mid R_j)$

These two conditions are supposed to capture the two dimensions that are relevant for content determination: the backward and forward probabilities. In particular, the first condition claims that, among all entities that increase the probability of R occurring, S_i is the one that increases this probability more. That is, the claim is that among all the stimulus eliciting R, S_i is the one that is more likely to produce R. This first condition is supposed to single out the stimulus that better correlates with the mental states. In contrast, the second condition compares different representational states. The idea is that R represents S_i only if R is the representational state that increases more the probability of S_i being the case. Here the probability that matters is the backward probability conditionalized on representational states.

New informational approaches such as Info have certain features that make them worth considering in detail. For one thing, they seem to solve the two most pressing problems of Dretske's approach, namely the problem of misrepresentation and its empirical implausibility. First of all, since these theories reject Dretske's suggestion that the likelihood of the referent given the representation has to be one, they make it possible for a state to represent S when S is not the case. Representational relations are grounded on statistical dependencies between entities, so in a given occasion a representational state might be caused by an entity that it is not in its extension. Secondly, new informational theories are also much more realistic than the previous proposals in this tradition. Indeed, as they argue, this approach might indeed capture the way neuroscientists reason (Usher, 2001, p. 320). For instance, following Hubel and Wiesel's (1959) methodology, many neuroscientists identify the referent of a neuronal structures in early vision with the stimulus that is more likely to elicit a stronger response. Along the same lines, an additional virtue of these approaches is that they provide a precise method for discovering the content of neural events. They make very determinate predictions about the content of representational states, which is extremely valuable in scientific projects (Eliasmith,2000, p. 71).

For these and other reasons, in recent years scientifically guided informational theories have been gaining prominence (e.g. Pezza and Terenzi, 2007; Rusanen and Lappi. 2012, Scarantino, 2015). In what follows, however, we will like to argue that this optimism is probably unfounded.

3 Representation and Metarepresentation

As we previously mentioned, SGIT are naturalistic theories of mental content, since they attempt to clarify the nature of the relation that holds between a representation and its object. This seems to require, at least, an answer to two questions: i) what is a representation and ii) what is the content of that representation. In this paper we will focus on the second question.[3] Accordingly, we will argue that SGIT fail to provide sufficient conditions for determining representational content. More precisely, we will show that SGIT lack the resources to distinguish between metarepresentations (which have another representation as its object) and representations that reliably correlate with another representation (but which do not have a representation as its object).

To develop our argument, in this paper we will focus on representations we have of our own mental states. These representations are interesting for several reasons. In the first place, we seem to, at least sometimes, know what we think, what we regret, what we perceive, what we fear, etc. These are particular instances of our general ability to represent our own mental representations. A satisfactory naturalistic theory of representation should be able to account for these metarepresentational states.

Secondly, understanding this metarepresentational capacity is not only interesting for its own sake. It is well-established that we usually attribute mental states to others in order to explain their behavior (that is what philosophers call 'folk psychology'). Furthermore, it is commonly held that a unique mechanism underlies mind-reading (attributing representations to others) and metacognition (attributing representations to oneself) and that both abilities are directly connected (cf. Nichols and Stich 2003). There is, however, a huge controversy on whether metacognition is prior to mindreading—that is, on whether the ability of mindreading depends on the mechanisms that evolved for metacognition—or the other way around. Defenders of the so called 'theory-theory' (Gazzinaga, 1995, 2000; Gopnik, 1993; Wilson, 2002) argue that when we mindread, we make use of a theory of human behavior known as 'folk psychology'. This theory, just like other folk theories such as folk physics, helps us to master our daily lives successfully. On this view, mindreading is essentially an exercise in theoretical reasoning. When we predict the behavior of others, for example, we make use of folk psychology and reason from representations of the target's past and present behavior and circumstances, to representations of the target's future behavior. For theory-theorists, if there is just one mechanism, then metacognition depends on mindreading: metacognition is merely the result of turning our mindreading capacities upon ourselves (for an excellent review of the evidence in favor of the claim that mindreading is prior to metacognition see Carruthers 2009, 2011). On the other hand, defenders of simulation theories of mind like Goldman (2006) suggest that metacognition is prior to mindreading. The attribution of mental states to others, on this view, depends upon our introspective access to our own mental states together with processes of inference and simulation of various sorts, where a simulation is the process of re-enacting or attempt to re-enact, other mental episodes. If metacognition is prior to mindreading, then the latter would also depend on the kind of metarepresentations we are considering. Recently, alternative approaches have also been developed, such as hybrid (Nichols and Stich, 2003) or minimalist theories (Bermudez, 2013).

Finally, the ability to represent our own mental states might also play an important role in consciousness. For example, David Rosenthal (1997, 2005) has defended that conscious states are those one is aware of oneself as being in. This transitivity principle motivates one of the most popular families of theories of consciousness: higher-order representational (HOR) theories.[4] HOR theories explain

[3] For a detailed discussion of whether Scientifically Guided Informational Theories can solve the first problem, see [authors].

[4] Defender of same-order theories (Kriegel, 2009, [author1]) agree with this idea. It is unclear whether defenders of such transitivity principle are committed to a representation of a representational state (cf. author1).

what it takes for states to be conscious by means of an awareness of that state. If such an awareness is to be unpacked as a form of representation (Kriegel 2009), then consciousness depends on metarepresentation. Although there is plenty of controversy on the nature of the higher-order representation—as on whether whether higher-order states are belief-like (Gennaro (1996, 2012), Rosenthal (1997, 2005)) or perception-like (Amstrong (1968), Carruthers (2003), Lycan (1996))—, HOR theories commonly claim that a conscious mental state is the object of a higher-order representation of some kind; i.e. on metarepresentation.[5]

So there are good reasons to postulate and investigate metarepresentations. As a result, a satisfactory theory of mental content should be able to explain what makes a representation the object of another representation. In the next section we want to argue that SGIT lack the resources to do so. In section 5, we will discuss whether this problem can be solved by endorsing a functional account that supplements (or substitutes) these interesting theories.

4 Scientifically-Guided Informational Theories and Metarepresentation

Can SGIT accommodate metarepresentations? The answer we will develop in this section is that probably not. Although for the sake of the argument we will grant that, in many cases, SGIT can account for the difference between *being caused by S* and *representing S* (and, in this way, solve a classical problem of previous informational theories such as Dretske's 1981) we will argue that they are unable to make this distinction in the context of metarepresentations. In other words, the central problem that SGIT faces is that of distinguishing a case in which a state R_1 *represents* another representational state R_2 from a case in which a representation $R_1^{'}$ represents some stimulus but it is regularly *caused* by another representational state $R_2^{'}$. Since $R_1^{'}$ and R_2 can correlate as good (or as bad) as $R_1^{'}$ and $R_2^{'}$, and correlations (conditional probabilities) are all the resources SGIT have to explain the differences, these cases pose a serious problem for SGIT. This is the main objection we will develop in this section.

As we said, in our articulation of the objection we will focus on a particular formulation of SGIT—Usher's proposal (although it important to keep in mind that our argument is supposed to apply much more broadly. See section 4.4). We will argue that INFO cannot distinguish metarepresentations from stimulus representations by considering the two conditionals. First, we will show that if INFO is used to establish that R_1 is a representation of a determinate stimulus, INFO could also be employed in order to show that R_1 is a metarepresentation of another mental state.[6] Secondly, we will argue that if R_1 is a metarepresentation, then the same theory entails that, under certain circumstances, it is rather a representation of an external stimulus.

[5] Rosenthal (2012) has recently defended that metacognition and the postulated higher-order representation has little in common beyond the fact that they both postulate higher-order psychological states. It should be noted that even if Rosenthal is right and consciousness does not require metacognition as it seems at least *prima facie*, defender of HOR theories still accept that consciousness depends on representation of our own mental states.

[6] Of course, INFO is incompatible with R representing the two states at the same time (since, *ex hypohtesi*, the conditions pick up a single state). The point is that the external stimulus is as good a candidate as the other mental state.

4.1 From representation to metarepresentation

Consider a red object moving toward a subject S, who is looking at it. S's brain will generate a visual representation R_{rm} in highly visual cortical areas. Given the widely accepted principle of functional specialization on which the visual system operates, we know that R_{rm} requires the existence of other representations. For instance, visual attributes like color and motion are processed by different systems (Livingstone and Hubel 1988, Zeki 1978, Zeki et al. 1991). Whereas color is processed mainly by the blobs of V1, the thin stripes of V2 and the V4-complex, motion is processed by a different pathway that goes from cells of layer 4B in V1 to the thick stripes in V2 and to V5 (**?**Shipp (1985), Sincich (2005), Zeki and Shipp (1988)). As a result, whenever we possess R_{rm} we also have two different representations: one of the color of the stimulus, call it 'R_r' and one of its motion, 'R_m'. Further processing in the visual system results somehow (Bartels and Zeki (2005), Milner (1974), Shadlen and Movshon (1999), Treisman and Gelade (1980)) into a representation that *binds* both features into a representation of a moving red object, R_{rm}.

According to INFO, R_{rm} represents a moving red object because:

1. Red moving objects are the most likely stimulus that produces R_{rm} [$P(R_{rm} \mid red\,moving\,object) > P(R_{rm} \mid S)$; for all $S \neq red\,moving\,object$]

2. R_{rm} is the representational state that increases more the probability of there being a red moving object. [$P(red\,moving\,object \mid R_{rm}) > P(red\,moving\,object \mid R_x)$; for any R_x of the subject such that $R_x \neq R_{rm}$.

For the sake of the argument, let us grant that INFO can satisfactorily exclude other stimuli from the content of the representation. The problem we would like to highlight is that in this scenario INFO will entail that R_{rm} is a metarepresentation: R_{rm} represents R_m.

First of all, condition 1 claims that a representational state represents whatever increases more its probability. Yet, at this point, problems begin. As R_m is part of the causal chain that leads to R_{rm}, we can hardly assume that the presence of a red moving objects increases more the probability of R_{rm} than the state that represents moving things (R_m) does; that is, it is far from obvious that $P(R_{rm} \mid red\,moving\,object) > P(R_{rm} \mid R_m)$. Given the structure of the visual system, the normal causal path leads from red moving things to R_m , which in turn leads to R_{rm}. And since moving objects cause R_{rm} by means of causing R_m, $P(R_{rm} \mid R_m)$ is going to be at least as high as $P(R_{rm} \mid red\,moving\,object)$; in other words, we cannot expect R_{rm} to carry more information about red moving object than the information it carries about R_m. Therefore, although cases in which there is a red moving object, R_{rm} is tokened and R_m does not occur are undoubtedly possible, we should expect them to be rare, especially in comparison with cases in which both $R_{rm}$and R_m are tokened, but there is no red moving object (something that happens, for instance, every time there is a red object that the system misrepresents as moving). The inequality $P(R_{rm} \mid red\,moving\,object) > P(R_{rm} \mid R_m)$ is satisfied just in case the former situation is more often than the latter, something that does not happen in ordinary conditions—although, as we will discuss in the next subsection, such odd conditions are possible, thereby preventing the possibility of metarepresentation. Thus, the correlation between the final representational state and red moving things should not be expected to be higher than the correlation between the former and the intermediate representation (actually we would expect quite the opposite!). Thus, condition 1 gives us no reason for thinking that R_{rm} represents a red moving object rather than the mental state R_m.

One might suggest that condition 2 can help to avoid this conclusion, but we think this is unlikely. As we saw, the second condition compares different representational states. It claims that R_{rm} represents red moving object because there is no other representational state R_x such that it is more probable that there is a red moving object when R_x is activated than when R_{rm} occurs. Now, since our strategy is to argue that it follows from INFO that R_{rm} represents R_m, we have to argue that there is no other representational state R_x such that it increases more the probability of R_m than R_{rm}. There might be many situations in which this might actually be the case. For instance, if most moving things are red, there will probably be no other representation R_x, such that $R_x \neq R_{rm}$ and $P(R_m \mid R_x) > P(R_m \mid R_{rm})$. For example, R_r, which represents red things will not do it, insofar as there are sufficient red things which do not move and so $P(R_m \mid R_r) < P(R_m \mid R_{rm})$. Thus, one can easily find counterexamples in which this second condition is also satisfied for R_m. Therefore, there are cases in which INFO will confuse a representation of a certain stimulus with a metarepresentation of an intermediate state . This suggests that INFO is an inadequate definition.

Before moving forward, let us briefly consider an objection to our argument. One might grant our point but insist that, *in general,* condition 2 guarantees that the intuitive result is delivered. For instance, in every environment in which most moving things are not red, it is not true that $P(R_m \mid R_{rm})$ is higher than $P(R_m \mid R_x)$ for all $R_x \neq R_{cm}$. In particular, if among the moving things there are more green than red items, then $P(R_m \mid R_{gm})$ is higher than $P(R_m \mid R_{rm})$, where $'R'_{gm}$ stands for a representation of green moving objects. Moreover, in this case $P(red\, moving\, object \mid R_{rm}) > P(red\, moving\, object \mid R_{gm})$, so apparently nothing would prevent R_{rm} from representing red moving things. Thus, can condition 2 at least help avoiding the conclusion that representations of stimuli are confused with metarepresentations in this restricted set of cases? Unfortunately, we think INFO is unlikely to be satisfying even in this restricted set of cases. We agree that if, as we just considered, most moving things are not red, then condition 2 blocks the possibility that R_{rm} represents R_m. However, in this scenario the problem simply reproduces for the representation of moving objects of the most common color. Suppose that such color is in fact green. Since *ex hypothesi* most moving things are green, condition 2 does not prevent R_{gm} from representing R_m. Avoiding this conclusion would require that there is another representational state, R_x, such that it is more probable that there is a green moving object when R_x is activated than when R_{gm} is activated. But none of the states involved in the cognitive process we are describing—neither the state that represents green, as we have previously seen, nor representations of other color moving object—will do. Sure, it is an open possibility that there is still some other mental state *not involved in the cognitive process* that correlates better with the intermediate representation thereby preventing this result. Nonetheless, whereas this might be the case in some particular case, it is unreasonable to believe that this is going to be the case for every single complex representation as SGITs would require. Likewise, since we cannot assume that $P(R_{gm} \mid green\, moving\, object) > P(R_{gm} \mid R_m)$, there is no reason for thinking that R_{gm} represents $green\, moving\, object$ rather than R_m.

Consequently, even in the restricted set of cases in which INFO can distinguish stimuli representations from metarepresentations due to a particular environmental structure, the same problem will simply reappear at a different location. The rejoinder, then, is probably unsuccessful.

4.2 From metarepresentation to representation

So far, the argument has intended to show that if INFO entails that R_1 is a representation of a certain stimulus, the same account could be used in order to show that R_1 is a metarepresentation of another mental state. Let us now try to argue for the converse claim, namely that, at least in some cases, if, according to INFO R_1 is a metarepresentation, then INFO implies it is a representation of an external stimulus.

Consider now a mental state that represents red things, R_r and a metarepresentational state, MR_r that has the former state as its object. Let us start discussing condition 2. It claims that MR_r is a metarepresentation of R_r only if MR_r is the representational state that increases more the probability of R_r. Here we have to show that this condition can also be satisfied with respect to an external object, i.e. there is also a stimulus S such that MR_r is the representational state that increases more its probability.

Consider, for instance, cases in which metarepresentations demand a higher degree of reliability than first-order representations. For example, at least in some circumstances, one might expect that the formation of a metarepresentation (like the belief that I am seeing something red) is more demanding in terms of reliability that what is required to actually have the first-order representation (i.e. to actually see red)t. In circumstances like that, MR_r might be the representational state that increases more the probability of a red object being there, because the tokening of the metacognitive state (MR_r) requires a higher threshold of reliability than the first-order representation (R_r). For illustration, consider a model according to which metacognition works as a Bayesian filter (Lau and Passingham (2006), Lau (2008)). In this case, MR_r is tokened only if the probability that the first order representation is tokened because it was caused by a red thing is higher than a certain threshold: if $P(R_r \mid red\,thing) > \theta$, being ϑ the threshold value. This might depend, for example, on the firing intensity of the neural network which serves as vehicle of representation, thereby avoiding noisy cases. Imagine that such threshold is set under certain circumstances to 0.8. This would mean that the activation of the metarepresentation requires a level of activation of the first-order representation ($R_{r-required}$) that happens with a conditional probability on the stimulus of 0.8 ($P(R_{r-required} \mid red\,object) > 0.8$): it is not enough that R_r is tokened but it has to be tokened and have certain intensity. On the other hand, all that is required in this respect for R_r to represent red things is that the conditional probability of the state relative to the stimulus is higher for red things than for any other stimuli. Imagine that the stimulus that more probably activates R_r which is not a red thing, is a pink thing, something that happens 15% of the time: $P(R_r \mid pink\,object) = 0.15$. If the red objects cause the activation of the neural structure more often than pink things—and *ex hypothesis* more often than any other stimulus—then R_r represents red things—at least insofar as condition 1 is regarded. Imagine that this happens 60% of the time: $P(R_r \mid red\,object) = 0.6$. In this case, $P(R_r \mid red\,object) > P(R_r \mid S)$; for all $S \neq red\,object$, which guarantees that condition 1 of INFO is satisfied. However, crucially, $P(R_r \mid red\,thing) = 0.6 < \theta = 0.8$, so the metarepresentation is more reliable than the first-order representation concerning the presence of a red object. Accordingly, in these circumstances $P(red\,object \mid MR_r) > P(red\,object \mid R_r)$, so MR_r would be the representational state that increases more the probability of red things.

Let's turn now to condition 1. MR_r is a metarepresentation of R_r only if R_r is the stimulus that is most likely to produce MR_r, i.e. $P(MR_r \mid R_r) > P(MR_r \mid R_x)$, for all $R_x \neq R_r$. To put this inequality into question we need to argue that if R_r is regularly caused by red stimuli, $P(MR_r \mid red thing)$ is at least as high as $P(MR_r \mid R_r)$. That would show that, if the first condition of INFO when applied to assess the content of MR_r is satisfied by R_r, there will probably a particular stimulus, red thing in our case, that also fulfills it.

However, as we argued in the previous subsection, this is hardly plausible. At least in ordinary circumstances, states tend to carry more information about their proximal causes than about their distal causes. The reason is quite simple indeed: the visual system sometimes makes mistakes. In some cases, R_r is tokened when there is no red thing around and in those cases the covariation between MR_r and red things also fails. However, in other cases R_r is tokened in the presence of a red thing and MR_r fails to be activated. Thus, we cannot expect MR_r to generally carry more information about red objects—the distal cause—than the one it carries about R_r—the proximal cause— and, as a result, the default assumption should be that $P(MR_r \mid R_r){>}P(MR_r \mid red thing)$. Ironically, the main problem of Dretske´s account (the possibility of misrepresentation) seems to come to the rescue of informational theories.

Unfortunately, however, the mere appeal to errors is unable to provide a satisfactory solution. In a nutshell, the problem of this suggestion we would like to highlight is that mistakes can also go in other directions. More precisely, the following three conditions might obtain: (1) MR_r is tokened, (2) there is a red moving thing and (3) there is no R_r. As a consequence, misrepresentation can decrease the correlation between MR_r and red moving things, but it can also decrease the correlation between MR_r and R_r. That show that in some circumstances it might be the case $P(MR_r \mid R_r) < P(MR_r \mid red thing)$. Thus, in these situations condition 1 cannot establish that MR_r is a metarepresentation of R_r and not a representation of red things. The following example might help illustrate the idea. Consider two different causal paths leading to the activation of MR_r. In the first one, a red thing causes the activation of R_r, which in turn activates under certain circumstances MR_r. Imagine that there is another stimulus, S, which can also cause the activation of MR_r. Call this second path 'the deviant path'. Clearly, MR_r does not represent S, because $P(MR_r \mid R_r) > P(MR_r \mid S)$)— this is why we call it 'deviant path'. Nonetheless, under certain plausible environmental conditions, this deviant path might cause certain troubles. In particular, imagine that there is a strong correlation between Ss and red things in the environment. In this circumstances, cases in which R_r misses its target—and hence it is not tokened despite there being a red object—might be cases in which nonetheless MR_r is tokened due to the deviant path. As a consequence, we would expect $P(MR_r \mid red thing) > P(MR_r \mid R_r)$). This is a simple example in which, according to INFO, MR_r would represent red things.

In reply, one might bite the bullet and claim, as the theory predicts, in these cases MR_r is not a metarepresentational state, but a first-order representation of red things. The problem with this suggestion is that the high correlation between S and red things is a contingent fact of a particular environment and, accordingly, it would be unreasonable to maintain that under such circumstances the organism fails to have the required metacognitive states. To make the point more pressing, suppose that the metacognitive state is that belief that I am seeing something red; if the previous argument is on the right track, Info would entail that there are environments in which I cannot form such a belief, because of a certain correlation of stimuli. Moreover, consider a HOR of consciousness, like the one described in section 3. According to it, undergoing a conscious experience depends upon a higher-order representation—that one is seeing red in the case of an experience as of red. Info when combined with a HOR theory of consciousness has the undesired consequence that in certain environments—one in which there is sufficiently high correlation between Ss and red things—the organism fails to have experiences as of red. This is extremely implausible.

At this point, a remark is required. Certainly, our arguments do not show that Info entails that *all* metarepresentational states actually represent distal stimuli. This should be obvious, since the arguments in this subsection assume a particular set of additional circumstances (the existence of a deviant path, etc...). Nonetheless, this fact does not diminish the force of our arguments. Info (and, in general, SGIT) seeks to provide general conditions for a mental state to possess a determined representational content. To support the view that these theories are unsuccessful, one need not show that it delivers the wrong results in *all* cases. The fact that it has unintuitive consequences in some clear cases and that it makes representational content to depend on certain features that seem irrelevant (such as the contingent correlation between S and red things in the case of deviant paths) should be enough for casting doubt on these approaches.

To sum up, it seems that in an important set of cases, if MR_r is a metarepresentation of R_r, then it will follow from info that MR_r is a representation of a red object. Furthermore, since in the previous section we have shown that the reverse conditional also holds, we conclude that INFO cannot adequately distinguish representations of external objects from metarepresentations.

4.3 A Rejoinder

Anticipating a similar objection, Eliasmith (2005a) remarks that "In general, statistical dependencies are too weak to properly underwrite a theory of content on their own.[...] because the highest dependency of any given vehicle is probably with another vehicle that transfers energy to it, not with something in the external world." (p. 1046). In an attempt to address this issue, he includes an additional condition that should allow Info to exclude other neuronal states as referents. In particular, he adds that the referent cannot "fall under the computational description", that is, there must not be any internal computational description relating the referent with the mental state such that it could account for the statistical dependence. Thus, according to him:

> The referent of a vehicle is the set of causes that has the highest statistical dependence with the neural responses under all stimulus conditions and *does not fall under the computational description.* (Eliasmith 2005a, p. 1047; Eliasmith 2000 p. 59-60; emphasis added)

Where the computational description refers "to the account of neural functioning provided by the theory of neural representation" (p.1047). For instance, activity in V1 has a high statistical dependence with activity in the thalamus, but the reason is that they are computationally related. With this additional clause, the latter can be ruled out as possible content.

Now, it is unclear to us what independent consideration can justify what seems to be a clear adhoc movement. But let us grant for the sake of the argument that there is some independent way of motivating this new condition. At first glance, one might think that it can solve the problem we were dealing with: despite the fact that a red moving object does not rise the probability of R_{rm} more than R_m, R_{rm} represents the former because R_{rm} falls under the computation description, since it is a component of the system. However, there are at least two compelling reasons why his proposal is unlikely to succeed.

First of all, note that computations are defined over representations. To know whether two causally related brain states are computationally related, one should know whether they are representations and how their content is related. Yet this is precisely what this condition is supposed to establish. The requirement that only entities that do not fall under the computational description can qualify as representational objects is of no use in a theory of representational content, because we need such a theory in order to determine which entities should be excluded. Put in a different way: a theory that presupposes the representational content of certain states cannot in turn be used to deliver these contents.

The second problem with this suggestion is that it seems to exclude too much, because we do indeed have some representations of our own neural states (which, arguably, also fall under a computational description). For instance, suppose that Higher-Order Representational (HOR) theories of consciousness are right and we need metarepresentations in order to have an experience as of red. In that case, if S is having an experience as of seeing red, she needs to have a metarepresentation of R_r, most probably in the dorsolateral prefrontal cortex (Lau and Passingham (2006), Lau and Rosenthal (2011)).[7]. Call this metarepresentation 'MR_r'. According to INFO, MR_r represents R_r because:

1. R_r is the most likely stimulus that produces MR_r [$P(MR_r \mid R_r) > P(MR_r \mid S)$; for all S distinct from R_r and MR_r]

2. MR_r is the representational state that increases more the probability of there being R_r [$P(R_r \mid MR_r) > P(R_r \mid R_x)$; for all R_x of the subject distinct from MR_r and R_r] .[8]

But note that, if Eliasmith's modification of INFO is accepted, this theory would be known to be false *a priori,* because it would be impossible for a state to represent another neuronal state in that way if both are computationally related. And although we think that the truth of HOR theories is far from established, it would be highly inadequate to exclude such a theory by the mere definition of what representing is. Consequently, we think that Eliasmith rejoinder is far from being fully satisfying.

4.4 Generalizing the argument

If the arguments so far have been on the right track, Ushers´s and Eliasmith´s SGIT lack the resources to allow us to say that R_{rm} represents a red moving object rather than R_m and, at the same time, that MR_{mr} represents R_{mr}. Moreover, the reasoning developed in the preceding sections suggests that this failure is rooted in the fact that they try to explain content by exclusively appealing to statistical dependence. Thus, *mutatis mutandis* one should expect the same problem to affect other SGIT that rely on correlations. For instance, consider Skyrms' theory (which, with slight modifications,

[7] cf. Bartels and Zeki (2005). According to them the binding of motion and color is a post-conscious process.

[8] Once metarepresentation enters into play, conditions 1 and 2 has to be slightly modified, for no state increases the probability of a state M more than M itself. Quantification is restricted accordingly in 1 and 2.

is also embraced by Birch, 2014). According to this approach, the informational content of a given representation R is a vector. More precisely, the informational content is a vector which tells us how a signal changes the probabilities of all states. If there are only four possible states of the world (S_1, S_2, S_3, S_4), the informational content of a signal should be calculated with the following formula: $< log_2 \frac{P(S_1|R)}{P(S_1)}, log_2 \frac{P(S_2|R)}{P(S_2)}, log_2 \frac{P(S_3|R)}{P(S_3)}, log_2 \frac{P(S_4|R)}{P(S_4)} >$. For example, in a given occasion the informational content of a certain signal could be $< 1.25, -\infty, -\infty, 0.68 >$ (the $-\infty$ components are going to end up with probability 0; this is just a side effect of using logarithms). In normal parlance, this signal tells you that the probability of S_1 and S_4 has been increased and that S_2 and S_3 are impossible. Thus, this signal represents $S_1 \vee S_4$, where the probability of S_1 being the case is higher than the probability of S_4.

Now, Skyrms does not provide a criterion for choosing the set of states, whose probabilities should be considered in the vector. For instance, do the probabilities of other mental states figure in the relevant vector? Depending on the answer he gives to this problem, Skyrms' approach seems to face a dilemma. If other mental states are excluded form the vector by definition, then the theory will face the same problem as Eliasmith's rejoinder, namely that of *a priori* excluding metarepresentations. If, on the other hand, the probabilities of other mental states are included in the vector, then representation of external stimuli and metarepresentations should be distinguished by their statistical dependencies, and we previously argued at length that this strategy will probably fail. In particular, we would expect a representation of the external world to have non-zero values for some external states and a metarepresentation to have non-zero values for some neuronal states. But, as we have seen, we have no reason to expect a difference (or, at the very least, a sufficiently significant difference) in the probabilistic vectors that correspond to, say, MR_{rm} and R_m. Consequently, if content is determined by conditional probabilities, we will have no way to distinguish them.

Likewise, other approaches like Rupert's (1999) or Scarantino's (2015) do not diverge from Usher's and Skyrms' theories in ways that would affect the main point of the paper. For instance, Rupert's account 1999 also analyzes representational relations in terms of probability relations between entities, although he only considers forward probabilities (i.e. conditionalized on entities) and restricts his account to representations of natural kinds. On this account, R represents a natural kind S iff members of S are more efficient in their causing R than are members of any other natural kind. However, the arguments we have presented concern entities that can plausibly qualify as natural kinds, so there is not reason for thinking his proposal can overcome the difficulties of other informational approaches.

Summing up, we think that the objections raised here probably generalize to many other Scientifically Guided Informational Theories. Although in previous sections we focused on Usher's informational theory, we think the problem is likely to affect any approach that seeks to define representational content in correlational terms.

5 Teleological Functions to the Rescue?

If our reasoning is correct, SGIT fail to provide a satisfactory account of representation, because they lack the resources for accommodating cases of metarepresentation. Even though we think that informational relations are likely to be an important element in our understanding how neural structures come to represent, an appeal to statistical dependencies between events is insufficient for providing a fully satisfactory naturalistic theory of content (see also Shea (forthcoming)). In this final section, we would like to explore some consequences.

Suppose the arguments developed in this essay are right. The first and most obvious solution is to complement SGIT with some other notion. But what else might be required? A popular suggestion is that metarepresentations and representations can be distinguished by appealing to the notion of function. The key idea, of course, is that metarepresentations are states whose *function* is to indicate other representational states, while other representations have the *function* to indicate external stimuli. Although there are different ways of spelling out the notion of function (Abrahams,2005; Cummins, 1975; Griffiths, 1993; Millikan, 1989; Mossio et al. 2009; Nanay, 2010), the standard (etiological) view has it that functions should be understood as selected effects, that is, as effects that were important for the selection of the trait. Thus, a particular brain structure (e.g. in the striate cortex) might have been selected for indicating external stimuli, while other structures (e.g. certain areas in the dorsolateral prefrontal cortex) might have been selected for indicating internal states of the organism. Indeed, there are already some proposals which try to combine informational and functional notions (Dretske, 1995; Lean, 2014; Martinez, 2013; Neander, 2013; Shea, 2007). So this is an interesting option that needs to be seriously taken into account.

Nonetheless, we would like to conclude by considering a risk. It might happen that adding the notion of function to an informational account has unexpected consequences for SGIT. More precisely, once functions are brought in, the notion of information might be shown to play no important role in the resulting naturalistic theory of content. Although a full discussion of whether information and functional notions can be coherently combined in that way lays beyond the scope of this essay, we would like to briefly argue why we think some tension might exist.

Suppose one is convinced by the arguments laid down in previous sections and accepts that carrying information is insufficient for delivering a satisfactory theory of content. As we just suggested, one could try to simply amend INFO by adding the notion of function. Accordingly, one could claim that the content of a given representational state is determined by the function to carry information about a certain state. That is, one could argue that the function of certain states is to correlate with certain state of affairs. Now, a difficulty with this idea is that the same problem we just saw with informational theories (i.e. that they lack the resources to establish whether a state is a representation of another representational state or the representation of an external stimulus), reappears at the level of function. After all, why should we think that the function of a representation is to carry information about an external stimulus rather than carrying information about another representational state? Just adding the notion of function might not be sufficient for a full answer to this worry.

Of course, this question could be addressed by specifying in more detail what is required for a state or a system to acquire a function. Perhaps an appeal to a specific aspect of the selection process or to the mechanism sending or receiving the signal could help with this problem. However (and this is the central point), if the notion of function can be made specific enough to solve the problem outlined here, it might happen that then the fact that a state has a high statistical dependence becomes largely irrelevant. While carrying information might still be an interesting property of certain states that might help explain why certain features of representational mechanisms evolved, carrying information would not constitute a necessary or a sufficient condition for a state to represent another state. Accordingly, on this approach the utility of the notion of information might be seriously called into question. Indeed, this result could jeopardize the scientific practices if—as we granted at the beginning—the implicit assumption that neuroscientists are making when establishing claims about the content of neuronal states is to be captured in informational terms.

Obviously, much more should be said in order to make this line of reasoning convincing. Nonetheless, we wanted to briefly call into question the assumption that information will utterly play a role in a satisfactory naturalistic theory of content; something that has not yet been established. At least, we have tried to show that information is unlikely to provide such a theory on its own.

References

Abrams, M.: 2005, Teleosemantics without natural selection, *Biology and Philosophy* **20**, 97–116.

Amstrong, D.: 1968, *A Materialist Theory of the Mind*, London: Routledge.

Bartels, A. and Zeki, S.: 2005, The temporal order of binding visual attributes, *Vision Research* **46**(14), 2280–2286.

Bermudez, J. L.: 2013, The domain of folk psychology, *in* A. O'Hear (ed.), *Minds and Persons*, Cambridge University Press.

Birch, J.: 2014, Propositional content in signalling systems, *Philosophical Studies* **171-3**, 493–512.

Carruthers, P.: 2003, *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge University Press.

Carruthers, P.: 2009, How we know our own minds: The relationship between mindreading and metacognition, *Behavioral and Brain Sciences* **32**(2), 121–138.

Carruthers, P.: 2011, *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford University Press.

Cummins, R.: 1975, Functional analysis, *Journal of Philosophy* **72**, 741–765.

Dretske, F.: 1981, *Knowledge and the Flow of Information*, Cambridge: MIT Press.

Dretske, F.: 1995, *Naturalizing the Mind*, The MIT Press.

Eliasmith, C.: 2000, *How neurons mean: A neurocomputational theory of representational content*, Unpublished Dissertation, Washington University in St. Louis.

Eliasmith, C.: 2003, Moving beyond metaphors: Understanding the mind for what it is, *Journal of Philosophy* **10**, 131–159.

Eliasmith, C.: 2005a, Neurosemantics an categories, *in* H. Cohen and C. Lafebvre (eds), *Handbook of Categorization in Cognitive Science*, Elsevier.

Eliasmith, C.: 2005b, A new perspective on representational problemss, *Journal of Cognitive Science* **6**, 97–123.

et al., J. L.: 2014, Getting the most out of shannon information, *Biology and Philosophy* **29**(3), 395–413.

Floridi, L.: 2010, *Information: A Very Short Introduction*, Oxford University Press.

Gazzaniga, M.: 1995, Consciousness and the cerebral hemispheres., *in* M. Gazzaniga (ed.), *The Cognitive Neurosciences*, MIT Press.

Gazzaniga, M.: 2000, Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition?, *Brain* **123**, 1293–1326.

Gennaro, R.: 2012, *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*, MIT Press.

Gennaro, R. J.: 1996, *Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*, John Benjamins.

Goldman, A. I.: 2006, *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, illustrated edition edn, Oxford University Press, USA.

Gopnik, A.: 1993, The illusion of first-person knowledge of intentionality, *Behavioral and Brain Sciences* **16**, 1–14.

Griffiths, P.: 1993, Functional analysis and proper functions, *British Journal for the Philosophy of Science* **44**(3), 409–422.

Hubel, D. H. and Wiesel, T. N.: 1959, Receptive fields of single neurones in the cat striate cortex, *Journal of Physiology* **148**, 574–59I.

Kriegel, U.: 2009, *Subjective Consciousness: A Self-Representational Theory*, Oxford University Press, USA.

Lau, H.: 2008, A higher-order bayesian decision theory of perceptual consciousness, *Progress in Brain Research* **168**.

Lau, H. and Passingham, R.: 2006, Relative blindsight in normal observers and the neural correlate of visual consciousness, *Proceedings of the National Academy of Science* .

Lau, H. and Rosenthal, D.: 2011, Empirical support for higher-order theories of conscious awareness, *Trends in Cognitive Sciences* **15**(8), 365–373.

Livingstone, M. S. and Hubel, D. H.: 1988, Segregation of form, color, movement, and depth: Anatomy, physiology, and perception., *Science* **240**, 740–749.

Lycan, W. G.: 1996, *Consciousness and Experience*, The MIT Press.

Martinez, M.: 2013, Teleosemantics and indeterminacy, *Dialectica* **67**(4), 427–453.

Millikan, R. G.: 1989, In Defense of Proper Functions, *Philosophy of Science* **56**(2), 288–302.

Milner, P.: 1974, A model for visual shape recognition, *Psychological Review* **81**(6), 521–535.

Mossio, M., Saborido, C. and Moreno, A.: 2009, An organizational account of biological functions, *British Journal for the Philosophy of Science* **60**(4), 813–841.

Nanay, B.: 2010, A modal theory of function, *Journal of Philosophy* **107**(8), 412–431.

Neander, K.: 2013, Toward an informational teleosemantics, *in* D. Ryder; J.Kingsbury; K. Williford (ed.), *Millikan and her critics*, Wiley-Blackwell.

Nichols, S. and Stich, S. P.: 2003, *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, illustrated edition edn, Oxford University Press, USA.

Pessa, E. and Terenzi, G.: 2007, Semiosis in cognitive systems: a neural approach to the problem of meaning, *Mind and Society* **6**, 189–209.

Rosenthal, D.: 2012, Higher-order awareness, misrepresentation and function, *Philosophical Transactions of the Royal Society of London* **367**, 1424–1438.

Rosenthal, D. M.: 1997, A theory of consciousness, *in* N. Block, O. J. Flanagan and G. Guzeldere (eds), *The Nature of Consciousness*, Mit Press.

Rosenthal, D. M.: 2005, *Consciousness and mind*, Oxford University Press.

Rupert, R.: 1999, The best test theory of extension: First principle(s), *Mind and Language* **14**(3), 321–355.

Rusanen, A. and Lappi, O.: 2012, An information semantic account of scientific models, *in* H. W. de Regt (ed.), *EPSA Philosophy of Science*, Springer, pp. 315–327.

Scarantino, A.: 2015, Information as a probabilistic difference maker, *Australasian Journal of Philosophy* .

Shadlen, M. and Movshon, J.: 1999, Synchrony unbound: a critical evaluation of the temporal binding hypothesis, *Neuron* **24**(1), 67–77.

Shannon, C.: 1948, A mathematical theory of communication, *The Bell System Technical Journal* **27**, 379–423.

Shea, N.: 2007, Consumers Need Information: Supplementing Teleosemantics with an Input Condition, *Philosophy and Phenomenological Research* **75**(2), 404–435.

Shea, N.: forthcoming, Neural signalling of probabilistic vectors, *Philosophy of Science* .

Shipp, S., . Z. S.: 1985, Segregation of pathways leading from area v2 to areas v4 and v5 of macaque monkey visual cortex., *Nature* **315**, 322–325.

Sincich, L. C., . H. J. C.: 2005, Input to v2 thin stripes arises from v1 cytochrome oxidase patches, *Journal of Neuroscience* **25**(44), 10087–10093.

Skyrms, B.: 2010, *Signals: Evolution, Learning, and Information*, Oxford University Press, Oxford.

Treisman, A. and Gelade, G.: 1980, A feature-integration theory of attention, *Cognitive Psychology* **12**, 97–136.

Usher, M.: 2001, A statistical referential theory of content: Using information theory to account for misrepresentation, *Mind and Language* **16**(3), 331–334.

Wilson, T.: 2002, *Strangers to Ourselves*, Harvard University Press.

Zeki, S. M.: 1978, Functional specialization in the visual cortex of the monkey., *Nature* **274**, 423–428.

Zeki, S. M., Watson, J. D. G., Lueck, C. J.and Friston, K. J. K. C. and Frackowiak, R. S. J.: 1991, A direct demonstration of functional specialization in human visual cortex., *Journal of Neuroscience* **11**, 641–649.

Zeki, S. and Shipp, S.: 1988, The functional logic of cortical connections., *Nature* **335**, 311–317.