Can Informational Theories account for Metarepresentation?

*M.A. Sebastián, M. Artiga**

(forthcoming in *Topoi*)

Abstract

In this essay we discuss recent attempts to analize the notion of representation, as it is employed in cognitive science, in purely informational terms. In particular, we argue that recent informational theories cannot accommodate the existence of metarepresentations. Since metarepresentations play a central role in the explanation of many cognitive abilities, this is a serious shortcoming of these proposals.

1 Introduction. Representation

Mainstream cognitive science maintains that our mind is a representational system; it holds 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. 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 rely heavily on the idea that information, understood as some form of statistical dependence, is the clue to understanding representations.

*This work is fully collaborative. Authors appear in random order. The final publication is available at link.springer.com

Of course, the claim that we can explain representations by appealing to some sort of information is not new, and can be traced back at least to Dretske (1981a). 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 neuroscientists when they claim, for instance, that certain neuronal activation in a particular 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 'Recent Informational Theories' (RITs). Some form or other of RIT has been defended for example by Usher (2001), Eliasmith (2000, 2003), Rupert (1999) or Skyrms (2010).

Interesting as RITs are, in this paper we argue that this kind of theories lack the resources to ground 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 that do not represent others —first-order representations—and representations that do—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, RITs are unable to do so.

The paper is organized as follows: section 2 presents RITs and section 3 clarifies the relevance of metarepresentations in our cognitive architecture. In section 4, we develop the idea that RITs are unable to account for the difference between metarepresentations and first-order representations and we consider some objections. Our argument intends to show that content cannot be fully determined solely in terms of statistical dependence relations. In section 5, we briefly discuss whether the notion of teleological function could help these approaches to solve this problem. We conclude by comparing our objections to classical arguments in the context of naturalistic theories of content.

2 Informational Theories

Recent Informational Theories (RITs) 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 to the classical problem of intentionality. The nature of representational phenomena would be finally understood.

The particular answer RITs give to this challenge connects with the long-standing informational tradition. The distinctive 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 (1981a), who tried to analyze representations (p. 160) and semantic content (p. 185) 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 $P(S \mid R) = 1$ (but, given background conditions alone, $P(S \mid R) < 1$) . 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 S represents (or carries the semantic content) that R only if it carries the information that R (Dretske, 1981b, p. 160,185) and there can be an informational relationship between two states only if both obtain (Dretske, 1981b, p. 65), 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 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, they 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 slightly

[1] We focus on theories of mental content that appeal to a probabilistic notion of information, Shannon's in particular (see Cover and Thomas, 2006, ch. 17 for a discussion of the relation between Fisher's [1925] notion of information, which is also probabilistic, and entropy as in Shannon's). 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 the way in which information is understood in naturalistic theories of content, a certain amount of correlation is sufficient for an entity to carry information about another entity (Floridi, 2010).

It might well be that there are other notions in the vicinity related to information, which are not probabilistic, like for example Kolmogorov's complexity [Kolmogorov, 1965]. However, there has not been any attempt, to the best of our knowledge, to show how they might relate to mental representation, and all theories of mental content that focus on information do so on a probabilistic notion. Indeed, we do not even begin to see how alternative notions would vindicate the practices in cognitive science or neurosciences.

[2] With respect to beliefs, Dretske (1981a, ch. 8) tried to address these problems by distinguishing a learning period (in which misrepresentation is still impossible) from a post-learning period, but it is generally agreed that this proposal can probably not solve any of these difficulties (Adams and Aizawa, 2010).

different ways. The main modification is to reject the requirement that $P(S \mid R) = 1$, which was the key assumption that made the theory unrealistic and rendered misrepresentation impossible. Yet dropping this assumption raises some questions. In particular, which probability should then be required for a state to represent another state? Any lower standard would seem arbitrary. 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 the probability of a certain state more than others. This is the central idea that has been developed in various ways by different authors.

Since a joint consideration of all RITs would be extremely complex, partly because they focus on different kinds of representations, 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 RITs (see section 4.4). More precisely, here we will concentrate on Usher (2001), because he defends an informational theory based on statistical dependence relations, which 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, 2005a, 2005b) 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 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 two sets 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 RITs is that representational content cannot just be determined by the fact that 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 combines two different conditions: (1) the mutual information R carries about S is greater than the information it carries about any other entity and (2) the mutual information between S and R is greater than the mutual 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 $j \neq i$

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 $j \neq i$[3]

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 stimuli 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 state. In contrast, the second condition compares different representational states. The idea is that R represents S_i only if R is the state that increases more than any other state within that organism 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 1, they make it possible for a state to represent S when S is not the case. Representational relations are now grounded on statistical dependencies between entities, so on a given occasion a representational state might be caused by an entity that 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 when they attribute representations

[3] Usher (2001) distinguishes an "external scheme"—characterized by condition 1—, and an "internal scheme"—characterized by condition 2. His position might be read as offering two different ways of independently fixing two distinct types of content and each condition as necessary and sufficient condition for its respective kind of content to be assigned. We think that INFO, which takes the conjunction of 1 and 2 as the necessary and sufficient conditions, probably captures better Usher's insights as well as the views of other philosophers such as Eliasmith's or Rupert's. In any case, we think that our reasoning does not depend on this interpretation. In section 4, we will argue that neither 1 nor 2 nor the conjunction of both can help our opponent to properly distinguish a representation of an external stimulus from a metarepresentation.

(Usher, 2001, p. 320). For instance, following Hubel and Wiesel's (1959) methodology, many neuroscientists identify the referent of a neuronal structure 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 RITs have been gaining prominence (e.g. Pezza and Terenzi, ?; Rusanen and Lappi. 2012, Scarantino, 2015). In what follows, however, we will argue that this optimism is probably unfounded.[4]

3 Representation and Metarepresentation

As we previously mentioned, RITs are naturalistic theories of mental content, since they attempt to clarify the nature of the relation that holds between a representation and its object.[5] 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.[6] Accordingly, we will argue that RITs fail to provide sufficient conditions for determining representational content. More precisely, we will show that RITs lack the resources to distinguish metarepresentations (which have another representation as their object) from first-order representations (which do not have a representation as their object) that reliably correlate with another representation.[7]

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, at least sometimes

[4] It is worth stressing that we will *not* be arguing against naturalistic theories of content in general (quite the opposite; we are sympathetic to this project). Thus, our goal here significantly differs from recent attempts to attack reductionist theories of representation, such as Ramsey (2007). Moreover, our argument is not intended to show that the notion of information should not play any important role in a theory of mental content. Rather, as we will make clear, we will argue against the idea that content attribution can be explain solely in terms of statistical dependencies as RITs commonly claim.

[5] In this paper we will assume that RITs seek to analyse the notion of 'representation' as it is employed in cognitive science (although some of the authors discussed here are more explicit about that goal than others). Nonetheless, it is worth pointing out that the relationship between folk, philosophical and scientific uses of 'representation' is far from straightforward (for a discussion, see Godfrey.Smith, 2006; Ramsey, 2007).

[6] For a detailed discussion of whether RITs can solve the first problem, see [authors].

[7] Similarly, in the same way that we assume that there is some independent way of specifying what representations are (i.e. we focus on semantic rather than metasemantic theories), we also wished to avoid the difficult question of how to exclude grue-like properties. For instance, why are we only considering properties like *being red*, *having a certain size* or *being a representation* R_i as candidate properties for being represented? Why should we exclude properties like *being grue*, *being red or a unicorn*? This is a very interesting question, but there are two reasons why we do not address it in this paper. First of all, this is a problem for any naturalistic theory of content, so it would be unfair to dismiss RIT for that reason (Sterelny, 1990). Secondly, the main goal of the paper is to show that, even if the reference class could be defined, RIT would have specific problems with metarepresentations.

we seem to 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 single 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 mindreading ability 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 attempting 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 (Bermúdez, 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.[8] HOR theories explain 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 higher-order states are belief-like (Gennaro, 1996, 2012; Rosenthal, 1997, 2005) or perception-like (Armstrong, 1968; Carruthers, 2000; 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.[9]

So there are good reasons to postulate and investigate metarepresentations. Furthermore, one would reasonably assume that the same kind of relation that holds between a first-order representation and its object—namely that of representation—also holds between a metarepresentation and the representation of it targets.[10] On this assumption, in the next section we want to argue that RITs lack the resources to distinguish, by means of content, a first-order representation and a metarepresentation. 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 Recent Informational Theories and Metarepresentation

Can RITs 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, RITs 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 a nutshell, the central problem that RITs face is that of distinguishing a case in which a state R_1 *represents* another representational state R_2 from a case in which a state $R_1^{'}$ represents some stimulus but it is regularly *caused* by another

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

[9] Rosenthal (2012) has recently defended that metacognition and the postulated higher-order representation have little in common beyond the fact that they both postulate higher-order psychological states. Nonetheless, even if Rosenthal is right and consciousness does not require metacognition, defenders of HOR theories still accept that consciousness depends on representation of our own mental states.

[10] Some of the authors of the theories we are considering explicitly restrict the scope of their theories to some kinds of representations (and remain silent, for instance, on whether they apply to both representations and metarepresentations). Nonetheless, it is hard to understand why one would provide a theory of content that works for one but not for the other. *Prima facie*, there is no reason to think that the relation that holds between the first-order representation and its target and between the metarepresentation and the first-order one is of a different kind, and we know of no argument to that effect. For this reason we will proceed on the assumption that both relations are of the same kind and hence that a common theory is required.

representational state $R_2^{'}$. Since $R_1^{'}$ and R_2 can correlate as well (or as badly) as $R_1^{'}$ and $R_2^{'}$, and correlations (conditional probabilities) are all the resources RITs have to explain the differences, these cases pose a serious problem for RITs. This is the main objection we will develop in this section.

As we have said, in our articulation of the objection we will focus on a particular formulation of RITs—Usher's proposal (although it is important to keep in mind that our argument is intended to apply much more broadly. See section 4.4). We will argue that INFO cannot distinguish metarepresentations from first-order representations by considering the two conditionals. First, we will show that if R_1 is a representation of a particular stimulus, INFO could be employed to show that R_1 is a metarepresentation of another mental state.[11] 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.

4.1 From representation to metarepresentation

Consider a red object moving towards a subject S, who is looking at it. S's brain will generate a visual representation of the red moving object, arguably in highly visual areas. Call such a representation 'R_{rm}'. 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, ?, ?]. 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 (Livingstone and Hubel 1987, ?, ?, ?). 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] in a representation that *binds* both features into a representation of a moving red object, R_{rm}.

In that case, we would intuitively attribute the content *moving red object* to R_{rm}. Indeed, this result seems to follow from INFO, since:

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

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

2. R_{rm} is the representational state that most increases the probability of there being a moving red object. [$P(moving\,red\,object \mid R_{rm}) > P(moving\,red\,object \mid R_x)$; 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 would actually entail that R_{rm} is a metarepresentation: R_{rm} represents another representational state.

First of all, condition 1 claims that a state represents whatever most increases its probability. However, problems begin at that point.[12] As R_m and R_r are part of the causal chain that leads to R_{rm}, we can hardly assume that the presence of a moving red objects increases more the probability of R_{rm} than the conjunction of the states that represents red things and moving things ($R_m \wedge R_r$) does; that is, it is far from obvious that $P(R_{rm} \mid moving\,red\,object) > P(R_{rm} \mid (R_m \wedge R_r))$. Given the structure of the visual system, the normal causal path leads from red moving things to R_m and R_r, which in turn leads to R_{rm}. And since moving objects cause R_{rm} by means of causing R_m and R_r, $P(R_{rm} \mid (R_m \wedge R_r))$ is going to be at least as high as $P(R_{rm} \mid moving\,red\,object)$; in other words, we cannot expect R_{rm} to carry more information about moving red object than the information it carries about R_m. Indeed, in the situation described, the $moving\,red\,object$, $(R_m \wedge R_r)$ *and* R_{rm} form a Markov chain in that order. A Markov chain is a particular kind of random process that undergoes transitions from one state to another on a state space characterized by the fact that the distribution of probabilities in transition from one state to the next depend *only* on the current state. It can be shown that if $X \rightarrow Y \rightarrow Z$ is a Markov chain then $MI(X;Y) \geqslant MI(X;Z)$ (this theorem of informational theory is known as the *data processing inequality* theorem. For a formal proof of the theorem see Cover and Thomas, 2006, pp.34-35). Therefore, as $moving\,red\,object \rightarrow (R_m \wedge R_r) \rightarrow R_{rm}$ is a Markov chain, $MI(moving\,red\,object; (R_m \wedge R_r)) \geqslant MI(moving\,red\,object; R_{rm})$; that is: $P(R_{rm} \mid moving\,red\,object) \leq P(R_{rm} \mid (R_m \wedge R_r))$.

Although cases in which there is a moving red object, R_{rm} is tokened, and R_m and R_r 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 and R_r are tokened, but there is no moving red object (something that happens, for instance, every time there is a red object that the system misrepresents as moving). The inequality $P(R_{rm} \mid moving\,red\,object) > P(R_{rm} \mid (R_m \wedge R_r))$ is satisfied just in case the former situation is more often than the latter, something that does not happen in ordinary conditions—although,

[12] Current articulation of this objection is deeply indebted to the comments of XX at YY.

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 moving red object rather than the conjoint state $R_m \& R_r$.[13] Therefore, intermediate first-order representations ($R_m \wedge R_r$) are likely to be better predictors of R_{rm} being tokened than any external stimuli. Condition 1 of INFO seems to be satisfied by these intermediate representations.

One might try to resist this reasoning by appealing to vision science and the well known phenomenon of perceptual constancy (Dretske, 1981a, p 157. See Burge, 2010 for an excellent review and discussion of the phenomenon).[14] It is widely accepted that whereas higher-level perceptual representations often correspond to stable properties of objects, early perceptual mechanisms usually respond to relatively local patterns of energy. The properties of objects we are sensitive to persist across changes in the energy reaching the senses (the proximal stimuli) and our visual system compensates for those changes facilitating our comprehension of the world. Classical examples of perceptual constancies include changes in size and in lightness. In many conditions, despite the fact that the viewing distance of an object (and, as a result, the projected retinal size) is significantly altered, its size is perceived as being the same. Similarly, the color of an object does not appear to vary when the object is viewed outdoors in the sunshine or indoors, in spite of the fact that there is a change of more than three orders of magnitude in the light intensity reflected from that object to the eyes (Garrigan and Kellman, 2008). In view of this situation, one might try to raise the following objection against our argument: consider a red object moving through three different lighting conditions—A, B and C. In each of these conditions different low level states would be activated ($R_m \wedge R_{rA}$, $R_m \wedge R_{rB}$, $R_m \wedge R_{rC}$), but they give rise to a unique higher-level perceptual representation, R_{rm}(for illustration purposes we can assume that three exhaust the conditions under which *normally* a moving red object gives rise to the activation R_{rm}). In this situation, none of the three possible low level states makes the tokening of R_{rm} more probable than a moving red object does, contrary to our reasoning. So it seems that the R_{rm} correlates better with redness than with any of the particular intermediate representations.

[13] Defenders of RITs might object to the *conjunctive* nature of the proposed metarepresentational content and claim that this is somehow illegitimate. However, note that the content they propose is also conjunctive, namely there is an obect that is red *and* moving. In any event, the principle that no mental state can represent more than one feature or state is highly dubious.

[14] We are grateful to XX for pressing us on this point.

Nonetheless, we think this reasoning is flawed. Even in cases of perceptual constancy there is a set of intermediate representations that better correlate with the activation of the higher-level state. The only difference with the previous example is that in this case one needs to include a disjunction of intermediate states. Consider color constancy: *ex hypothesi,* we know that R_{rm} requires the activation of R_{rA} or R_{rB} or R_{rC}. Thus, if in the previous examples, INFO entails that R_{rm} represents $R_m \wedge R_r$, in the case of perceptual constancy INFO implies that R_{rm} represents $R_m \wedge (R_{rA} \vee R_{rB} \vee R_{rC})$. Since $P(R_{rm} \mid R_m \wedge (R_{rA} \vee R_{rB} \vee R_{rC})) \geq P(R_{rm} \mid \textit{moving red object})$ the same reasoning we developed earlier can also be employed here. At this point, it's worth mentioning that Dretske (1981a, p. 158) thought that perceptual constancies could offer a solution because he assumed that $P(\textit{Distal} \mid \textit{Mental State}) = 1$ (and the reason proximal states are not represented is that $P(\textit{Proximal} \mid \textit{Mental State}) < 1$) . However, the fundamental assumption of RITs is that one cannot presuppose this strong probability, so it is hard to see why it is supposes that this strategy should work in the first place.

Let's consider now condition 2. One might hope that this requirement can help to avoid the conclusion that R_{rm} is a metarepresentation, but a closer look shows that this is an unlikely result. As we saw, the second condition compares different representational states. It claims that R_{rm} represents a moving red object because there is no other representational state R_x such that it is more probable that there is a moving red 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 $R_m \wedge R_r$, we only need to show that there are circumstances under which there is no other representational state R_x that increases more the probability of $R_m \wedge R_r$ than R_{rm}. It is easy to find such situations in which this might actually be the case. For instance, if R_{rm}, R_r and R_m were the only representations in a particular cognitive system, then, trivially, there would not be such a R_x. The same conclusion follows, for example, if most moving objects were red and most red objects were moving. In this case, there will probably be no other representation, R_x different from R_{rm} such that $P(R_m \wedge R_r \mid R_x) > P(R_m \wedge R_r \mid R_{rm})$.

Therefore, many first-order representational states would be wrongly classified as metarepresentations by INF.[15]

[15] Two clarifications are in order at this point.

First, one might note that the difficulty discussed so far seems to be just a particular instance of the well-known distality problem that causal theories face. And she might not be completely wrong. However, the problem metarepresentations pose to RITs goes beyond this, as we are about to see in the next section. In the conclusion we will argue in detail why the difficulty pointed out in this paper is more general and profound than the classical distality problem.

Second, according to some views, one state representing another is not sufficient for metarepresentation unless the former represents the later *as* a representation (see, for instance, Shea, 2014). Thus, one might object that although the argument presented here shows that R_{rm} represents R_m rather that a red moving object, this is not enough for counting as a metarepresentation. In reply it should be noted that in order to provide an explanation of what it takes

4.2 From metarepresentation to representation

So far, the argument has intended to show that if R_1 is a representation of a certain stimulus, INFO can often be used to show that R_1 is in fact 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. If R_r represents a red object, then MR_r represents that R_r is tokened, something that happens for example when we entertain a thought that we are seeing something red or when we are undergoing an experience as of red if higher order theories of consciousness are correct. Let us analyze the prediction that INFO would made in different circumstances.

Let us start with 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 indeed be satisfied by an external object, i.e. there is also a stimulus S such that MR_r is the representational state that increases more its probability. An example is provided by 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). 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). There are two properties of this model that are relevant. On the one hand, 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. θ might depend, for example, on the optimal way of avoiding noise in the firing intensity of the neural network which serves as vehicle of representation. Suppose that such a 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

to represent something as a representation we need to answer the question of what distinguishes representational states from other states. RITs are not intended to offer such an answer (and arguably they lack the resources to do it. For discussion see author, year). In any case, the main point we are making here is that according to the theories under consideration R_{rm} does not represent a red moving object, as it should do.

of 0.8 ($P(R_{r-required} \mid red\,object) \geq 0.8$)): it is not enough that R_r is tokened but it also needs to 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 pink objects cause R_r 15% of time and red objects 60% (14% of the time R_r is caused by something that is neither a red nor a pink object), i.e. $P(R_r \mid pink\,object) = 0.15$ and $P(R_r \mid red\,object) = 0.6$. In this scenario, $P(R_r \mid red\,object) > P(R_r \mid S)$; for all $S \neq red\,object$. which guarantees that at least condition 1 of INFO for R_r to represent red object is satisfied. Nevertheless, 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 condition 2 is satisfied by MR_r and $red\,object$ (and not by MR_r and R_r). 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. $P(MR_r \mid R_r) > P(MR_r \mid S_x)$, for all $S_x \neq R_r$. 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 be a particular stimulus, red thing in our case, that also fulfills it. Unfortunately, given what we argued in the previous subsection, finding a counterexample to this condition is extremely difficult. 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. Sometimes R_r is tokened when there is no red thing around and in those circumstances 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 generally to 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 when one is trying to argue that metarepresentational states are wrongly classified as first-order representations.

Nonetheless, although we agree that in general metarepresentations would satisfy condition 1, we think it is also possible to find some counterexamples. More precisely, a counterexample would need to

satisfy the three following requirements: (1) MR_r is tokened, (2) there is a red thing and (3) there is no R_r. This scenario would reduce the correlation between MR_r and R_r without affecting the correlation between MR_r and red things, so it will show it is possible that $P(MR_r \mid R_r) < P(MR_r \mid redthing)$. What we need is a case in which $redthing$, R_r and MR_r do not form a Markov chain and hence the *data processing inequality* theorem does not apply. Consider, for instance, 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. Let us suppose 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 problems. In particular, imagine that there is a strong correlation between Ss and red things in the environment. In these circumstances, cases in which R_r misses its target—and hence 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 $P(MR_r \mid redthing) > P(MR_r \mid R_r)$). This is a simple example in which, according to INFO, MR_r would represent red things.

At this point, a caveat is important. Note that 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 their force. INFO (and, in general, RITs) seeks to provide general conditions for a mental state to possess a determined representational content. To argue that these theories are unsuccessful, one need not show that they deliver the wrong results in *all* cases. The fact that they have unintuitive consequences in some clear circumstances and that they make representational content 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 one of the objections that we have presented, Eliasmith (2005b) 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, 2005b, 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 as regards 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, *prima facie* this move seems to be *ad hoc*. It is unclear to us what independent considerations can justify this claim. No principled reason is provided for restricting the scope of the theory other than the fact that it fails to accommodate certain cases.

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 could solve the problem we were dealing with: despite the fact that a moving red object does not increase the probability of R_{rm} more than $R_m \wedge R_r$, R_{rm} represents the former because there is a computational description of the visual system under which both $R_m \wedge R_r$ and R_{rm} fall. However, there are at least two compelling reasons why Eliasmith's proposal is unlikely to succeed.

First of all, some have argued that (at least some) computations are individuated by appealing to representations (Aydede [2005], Burge [2010], Peacocke [1999], Rescorla [2012], Sprevak [2010], Shagrir [2001]). This view is controversial (Piccinini [2008], Fresco [2010]; for a discussion, Fodor [1987, 1994], Shea [2013], Sprevak [2010]), but if it were correct, the rejoinder would be in trouble. If computations are defined over representations, to know whether two causally connected brain states

are computationally related, one should have to ascertain in advance 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. In other words: a theory that presupposes that certain brain states are representations with such and such content 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.[16] In that case, if a subject 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).[17] 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].[18]

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's rejoinder is far from being fully satisfying.

In a footnote Usher (2001, p. 326) hints at an idea that one could use as a reply to the concerns raised here.[19] As a reply to a Fodorian objection, Usher seems to suggest that a naturalistic theory of concepts should be restricted to what he calls 'objective world properties', which exclude 'subject dependent properties'. Could one avoid the worries pointed out in this paper by saying that INFO only

[16] Cases like blindsight [Humphrey and Weiskrantz, 1967, Humphrey, 1974, Weiskrantz, 1986] seem to suggest that there are visual representations in the absence of conscious experience.
[17] cf. Barteks et al. (2005). According to them the binding of motion and color is a post-conscious process.
[18] Once metarepresentation enters into play, conditions 1 and 2 have 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.
[19] We want to thank a reviewer for pressing us on this issue.

works with respect to object dependent properties? We doubt that a response along these lines can be satisfactory. First of all, as we argued earlier, adding this condition just because it allows the theory to address a potential worry seems to be *ad hoc*. Furthermore, at first glance there is no theoretical reason for thinking that we need two naturalistic theories of content, one for the representation of objective world properties and the other for subjective dependent properties. Secondly, the notion of 'subject dependent property' is not explained in any detail, but a natural interpretation suggests that it includes (at least) mental properties. If that is true, then this this rejoinder would show that INFO cannot accommodate representations of one's mental states as well as the mental states of other people. Indeed, if the notion of 'subject dependent property' is understood as usual, INFO might not work for the representations of colors or money or, for instance. Thus, it seems to throw the baby out with the bathwater. Consequently, we think that this rejoinder is unlikely to succeed.

4.4 Generalizing the argument

If the arguments so far have been on the right track, in certain cases Usher's and Eliasmith's RITs lack the resources to allow us to say that R_{rm} represents a moving red object rather than $R_m \wedge R_r$ and, at the same time, that MR_{mr} represents R_{rm}. Moreover, the reasoning developed in the preceding sections suggests that this failure is rooted in the fact that they try to explain content by appealing exclusively to statistical dependence. Thus *mutatis mutandis* one should expect the same problem to affect other RITs that rely on correlations. For instance, consider Skyrms (2010)' theory (which, with slight modifications, is also embraced by Birch, 2014). He develops his account within a game-theoretic framework, but one could suggest extending it to the content of brain states.[20] On this view, the informational content of a given representation R would be 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

[20] One might reasonably wonder whether Skyrms' theory could be applied to cognitive systems, such as the ones we are interested in; after all, he accepts multiple contents and he works within a sender-receiver framework that does not easily fit with cognitive systems. We agree that these are questions to be addressed by anyone interested in using Skyrms' approach in the context of neuroscience (for an interesting arguments in this direction see Cao, 2012), but those are orthogonal to the question addressed here. For the sake of the discussion, we will assume that those problems can be worked out.

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 question, Skyrms' approach seems to face a dilemma. If other mental states are excluded from the vector by definition, then the theory will share the problem of Eliasmith's rejoinder, namely that of excluding metarepresentations *a priori*. 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) do not diverge from Usher's and Eliasmith' theories in ways that would affect the main point of this paper. Rupert's 1999 account 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 objections we have presented concern entities that can plausibly qualify as natural kinds, so there is no reason to suppose that his proposal can overcome the difficulties of other informational approaches.

Summing up, we think that the objections raised here probably generalize to many other RITs. 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, which is the distinctive assumption of RITs.

5 Teleological Functions to the Rescue?

If our reasoning is correct, RITs fail to provide a satisfactory account of representation. Even though we think that informational relations are likely to be an important element in our understanding of 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 RITs with some other notion. But what else might be required? Dennett (who we think would be sympathetic to the results of this paper) describes semantic information as 'design worth getting' (Dennett, 2017, p. 115). Since talking of an item's 'design' is another way of referring to its function, Dennett's idea connects with the classical thread that tries to partly define representations in functional terms (Millikan, 1984; Papineau, 1993; Neander,1995; Godfrey-Smith, 1996). Thus, according to this line of reasoning, one should be able to distinguish metarepresentations and representations 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 taken seriously 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 RITs. More precisely, once functions are brought in, the notion of information might be shown to play no substantive 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 lies beyond the scope of this essay, we would like to briefly sketch the reason why we think some tension might exist.

Suppose one holds that representational content is determined by both functional and informational relations: a difficulty with this idea is that the same problem we just raised against 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 about another representational state? Just adding the notion of function might not be sufficient for a full answer to this worry (see Neander, 1995, 2013). 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 key point), if the notion of function can be made specific enough to solve the problem outlined here, then the fact that a state has a high statistical dependence might become largely irrelevant. While carrying information might still be an interesting property of certain states, it would not constitute a necessary or a sufficient condition for a state to represent another state. Accordingly, some ways of developing this idea might call into question the utility of the notion of information.

Obviously, much more should be said in order to make this line of reasoning compelling. The aim of this section, however, was much more modest. We just wanted to bring two ideas to the fore. First, that the notion of function is a promising tool for solving the problems highlighted in this paper. Second, that there might be some tension between a functional and an informational theory of content. Whether this tension can be dissolved is an open question.

6 Concluding Remarks: The Metarepresentational Challenge vs. Classical Problems

Recent Informational Theories of content have been gaining prominence in the philosophical and scientific literature. Assessing their merits is important because they seem to capture the scientific practice of content attributions within cognitive sciences. In this paper, we have tried to identify an underlying problem shared by all these approaches, namely that they fail to provide the tools for grounding the distinction between metarepresentations and first-order representations. Given the central role that these concepts play in current cognitive theories, we think that this is an important drawback that has not been sufficiently appreciated in the literature.

It is worth stressing that the metarepresentational challenge is an original problem. Some might be tempted to identify the objection presented here with the classical problem of indeterminacy (Fodor, 1990); nonetheless, this thought would be highly misleading. A theory suffers from the classical problem of indeterminacy if it lacks the resources to univocally determine which among some plausible candidates is the content of a mental state (see also Neander, 1995; Martinez, 2013). Purely informa-

tional theories might fall prey to this problem as well, but since no naturalistic theory has provided an uncontroversial solution to this issue, so it would be unfair to reject RITs for that reason. Yet the problem we are suggesting is *not* that the content of representational states is not determinate enough. In each of the examples we considered, we identified the specific entity that would qualify as the content of a certain representational state. Our objection is that the determined content entailed by the theory is the *wrong* one.

Alternatively, some might be tempted to identify the objection presented here with the *distality problem*; a common problem for theories that rely on the idea that content is a matter of indication or reliable causes. In a nutshell the problem is the following: consider a simple system that detects the presence of red objects in the environment. In this case, if an inner state indicates the presence of red objects it will also indicate the more proximal feature—say the presence of the light approaching the receptors—, as well as the more distal feature in the causal chain—like, for example, the presence of licopene in the detected object. In fact, the problem presented in 4.1 is linked to this one: any theory that relies solely on statistical dependence for content attribution will tend to relate content with a proximal state given the *data processing inequality.* If our argument were restricted to the one offered in 4.1 then one might reasonably think that this is a refined version of the distality problem, understood in the latter sense: the content involves a proximal state when it should concern a distal one, the external stimuli. However, we go beyond this idea, since we do not restrict our reasoning to one direction along the distality line. In 4.2, we show, that the theory predicts that sometimes the content relates to the more distal cause when it should be to the more proximal one—to the best of our knowledge, there is no previous argument in the literature against any naturalistic theory of mental content in this direction. The difficulty is therefore more general: the argument is not that content is always proximal; the objection is that content is often at the wrong level. Moreover, the problem that metacognition poses is a deeper difficulty, since it is specially relevant and worrisome for the explanation of behavior in comparison with other cases in which the content attribution fails to be the intuitive one. The reason is that in these examples it is very hard for the proponent of IT to simply bite the bullet and accept that the content of the mental state is the one predicted by the theory—*pace* the intuition.[21] Consider for illustration the infamous case of the frog and its alleged representation of a fly. A theory falls pray of the distality problem, in this latter sense, when it delivers a proximal

[21] To motivate this move one could think of a similar case in which the theory delivers counterintuitive results and the bullet is bitten like the case of Millikan's (2000) reply to Pietroski's (1992) thought experiment, where she accepts that the kimu's states represent *fewer snorfs this way*, rather than something like *red this way* (see also Price, 2001).

cause, say the presence of the light approaching the receptors, as the content of the state, when the intuitive result is the distal one: the presence of the fly. In response, one might very well attempt to bite the bullet, accept that the proximal cause is the content of the state, and explain the behavior of the frog by means of such representation. However, in the metacognitive case biting the bullet is extremely implausible because the behavior that can be explained is radically different depending on whether we postulate a first-order representation or a second-order representation and the cognitive capacities associated with them.

We have addressed the most recent informational theories of content, and shown in some detail why they face this important difficulty. We can conclude that RITs—but maybe not other naturalistic theories of mental content—are unable to ground the distinction between first-order and second-order representation that is so important for cognitive science. Cognitive theories make use of a clear distinction between representations of external stimuli and representations of representations that a satisfactory theory of mental content should accommodate.

Finally, if the arguments presented here are sound, they will have important consequences for cognitive science. If, as we granted at the beginning, the implicit assumption that neuroscientists often make when establishing claims about the content of neuronal states is to be captured in informational terms, then this result could jeopardize some scientific practices. A full examination of this question, however, is work for another paper.

References

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

F. Adams and K. Aizawa. Causal theories of mental content. *Stanford Encyclopedia of Philosophy*, 2010. URL `http://plato.stanford.edu/entries/content-causal/`.

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

Murat Aydede. Computationalism and functionalism: Syntactic theory of mind revisited. In G. Irzik and G. Guezeldere, editors, *Turkish Studies in the History and Philosophy of Science*. Springer, Dordrecht, 2005.

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

J. L. Bermudez. The domain of folk psychology. In A. O'Hear, editor, *Minds and Persons*. Cambridge University Press, 2013.

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

T. Burge. *The Origins of Objectivity*. Oxford University Press, 2010.

Rosa Cao. A teleosemantic approach to information in the brain. *Biology and Philosophy*, 1:49–71, 2012.

P. Carruthers. *Phenomenal Consciousness: a naturalistic theory*. Cambridge: Cambridge University Press., 2000.

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

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

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

D. Dennett. *From Bacteria to Bach and Back*. Norton and Company, 2017.

F. Dretske. *Knowledge and the Flow of Information*. The MIT Press, 1981a.

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

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

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

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

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

C. Eliasmith. Neurosemantics an categories. In H. Cohen and C. Lafebvre, editors, *Handbook of Categorization in Cognitive Science*. Elsevier, 2005b.

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

R.A. Fisher. Theory of statistical estimation. *Proceedings Cambridge Philosophical Society*, 20(5): 700–725, 1925.

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

J. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, 1987.

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

Jerry Fodor. *The Elm and the Expert*. MIT Press, Cambridge, 1994.

N. Fresco. Explaining computation without semantics: Keeping it simple. *Minds and Machines*, 20 (2):165–181, 2010.

Patrick Garrigan and Philip J. Kellman. Perceptual learning depends on perceptual constancy. *Proceedings of the National Academy of Sciences*, 5(6):2248–2253, 2008.

M. Gazzaniga. Consciousness and the cerebral hemispheres. In M. Gazzaniga, editor, *The Cognitive Neurosciences*. MIT Press, 1995.

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

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

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

P. Godfrey-Smith. *Complexity and the Function of Mind in Nature*. Cambridge University Press, 1996.

P. Godfrey-Smith. Mental representation, naturalism and teleosemantics. In MacDonald and D. Papineau, editors, *Teleosemantics*. Oxford University Press, 2006.

Alvin I. Goldman. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press, USA, illustrated edition edition, July 2006. ISBN 0195138929.

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

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

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

Nicholas Humphrey. Vision in a monkey without striate cortex: a case study. *Perception*, 3:241–255, 1974.

Nicholas Humphrey and L. Weiskrantz. Vision in monkeys after removal of the striate cortex. *Nature*, 215(5101):595–597, 1967.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1(1):1–7, 1965.

Uriah Kriegel. *Subjective Consciousness: A Self-Representational Theory*. Oxford University Press, USA, October 2009. ISBN 0199570353.

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

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

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

M. S. Livingstone and D. H. Hubel. Connections between layer 4b of area 17 and the thick cytochrome oxidase stripes of area 18 in the squirrel monkey. *Journal of Neuroscience*, 7(11):3371–3377, 1987.

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

William G. Lycan. *Consciousness and Experience*. The MIT Press, September 1996. ISBN 0262121972.

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

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

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

Ruth Millikan. *On Clear and Confused Ideas: An Essay about Substance Concepts.* Cambridge: Cambridge University Press, 2000.

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

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

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

K. Neander. Misrepresenting & Malfunctioning. *Philosophical Studies*, 79:109–141, 1995.

K. Neander. Toward an informational teleosemantics. In D. Ryder; J.Kingsbury; K. Williford, editor, *Millikan and her critics.* Wiley-Blackwell, 2013.

Shaun Nichols and Stephen P. Stich. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford University Press, USA, illustrated edition edition, October 2003. ISBN 0198236107.

D. Papineau. *Philosophical Naturalism.* Basil Blackwell, 1993.

Christopher Peacocke. Computation as involving content: A response to egan. *Mind and Language*, 9:195–202, 1999.

G. Piccinini. Computation without representation. *Philosophical Studies*, 137(2):205–241, 2008.

P. Pietroski. Intentionality and Teleological Error. *Pacific Philosophical Quarterly*, 73:267–282, 1992.

Carolyn Price. *Functions in Mind: A Theory of Intentional Content.* Oxford University Press, Oxford, 2001.

W. Ramsey. *Representation Reconsidered.* Cambridge University Press, 2007.

Michael Rescorla. How to integrate representation into computational modeling, and why we should. *Journal of Cognitive Science*, 13:1–38, 2012.

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

David M Rosenthal. A theory of consciousness. In Ned Block, Owen J Flanagan, and Guven Guzeldere, editors, *The Nature of Consciousness*. Mit Press, 1997.

David M. Rosenthal. *Consciousness and mind*. Oxford University Press, 2005. ISBN 9780198236962.

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

A. Rusanen and O. Lappi. An information semantic account of scientific models. In Henk W. de Regt, editor, *EPSA Philosophy of Science*, pages 315–327. Springer, 2012.

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

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

Oron Shagrir. Content, computation, and externalism. *Mind*, 110:369–400, 2001.

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

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

N. Shea. Reward prediction error signals are meta-representational. *Nous*, 2(48):314–341, 2014.

Nicholas Shea. Naturalizing representational content. *Philosophy Compass*, 8:496–509, 2013.

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

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

M. Sprevak. Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science*, 41:260–270, 2010.

K. Sterelny. *The Representational Theory of Mind: An Introduction*. Oxford University Press, 1990.

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

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

L. Weiskrantz. *Blindsight: A Case Study and Implications*. Oxford University Press USA, 1 edition, November 1986. ISBN 0198521294.

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