Rev.Phil.Psych. (2020) 11:613–627
https://doi.org/10.1007/s13164-018-0408-1

Informational Theories of Content and Mental Representation

Marc Artiga[1] **· Miguel Ángel Sebastián**[2]

Published online: 11 July 2018

Abstract Informational theories of semantic content have been recently gaining prominence in the debate on the notion of mental representation. In this paper we examine new-wave informational theories which have a special focus on cognitive science. In particular, we argue that these theories face four important difficulties: they do not fully solve the problem of error, fall prey to the wrong distality attribution problem, have serious difficulties accounting for ambiguous and redundant representations and fail to deliver a metasemantic theory of representation. Furthermore, we argue that these difficulties derive from their exclusive reliance on the notion of information, so we suggest that pure informational accounts should be complemented with (or perhaps substituted by) functional approaches.

Keywords

1 Introduction

Representations are puzzling entities. More than fifty years after the cognitive revolution and at a time in which representations are widely attributed in cognitive science, fundamental questions about their nature still remain unanswered. For one thing, it is not obvious what makes it the case that certain states are representations and some are not. For another, we lack a satisfactory account of what determines representational

✉ Marc Artiga
marc.artiga@uv.es

[1] Universitat de València, València, Spain

[2] IIF- Universidad Nacional Autónoma de México, Ciudad de México, Mexico

content. Now, given that cognitive scientist systematically attribute representations, it is not unreasonable to suppose that they might be implicitly assuming a set of intuitive conditions that are sufficient or even necessary for a state to qualify as a representation. Following this intuition, recently some philosophers and psychologists have tried to unravel this intuitive methodology and develop it into a full-blown naturalistic theory of representation. Since the notion of information plays an essential role, we will call them 'Scientifically Guided Informational Theories' ('SGIT', for short).

In this essay we would like to critically assess SGIT. In a nutshell, we will argue that, even if some SGIT might capture central assumptions in current scientific practice, they fail to satisfactorily explain the nature of representations and representational content. More precisely, the four objections we will develop is that SGIT do not account for some cases of error, fall prey to the wrong distality attribution problem, have serious difficulties accounting for ambiguous and redundant representations and fail to deliver a metasemantic theory of representation. Hence, although these accounts might faithfully embody (at least some of) the intuitive strategies employed in neuroscience in order to attribute representations, we will argue that they fail as theories of representation.

2 Representation and Information

Informational theories have a long history. One of the first and better known informational theories was Dretske's (1981a), 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 $P(S \mid R) = 1$, but $P(S) < 1$ given background conditions alone. While the idea of explaining semantic properties in terms of information was revolutionary and very influential, there were at least two deep problems with Dretske's proposal that caused a continuous loss of support for informational accounts. First of all, in the natural world it is extremely difficult (if not impossible) to find two different states such that the existence of one of them makes the other certain (even if certain background conditions are assumed). Secondly, it was incompatible with the fact that representational states can misrepresent. On Dretske's approach, a brain state can represent a state of affairs only if both obtain, so a typical case of misrepresentation (which usually involves an existing state representing a non-existing one) is rendered impossible.[1] These and other difficulties led most people to think that a purely informational theory of content was unworkable.

Recently, however, informational theories are reviving. To avoid these problems, SGIT define and use the notion of information in different ways. The common thread

[1] Of course, Dretske (1981a) was well aware of these problems, and he tried to solve them by distinguishing a learning period (in which misrepresentation is impossible) from a post-learning period. Unfortunately, it is widely agreed that this proposal still faces daunting problems. For one thing, it seems that the same difficulties reappear at the learning period.

is the rejection of the requirement that $P(S \mid R) = 1$, which was the origin of the two main objections to Dretske's approach. Instead, many of them rely on the comparison between probabilities. According to SGIT what is relevant is not how much a signal raises the probability of another state, but whether it raises the probability of another state more than any other representation does. In other words, these views focus on the distinctive statistical dependence between a representation and its referent.

Eliasmith (2000, 2005a, b, 2013), for instance, has defended an informational theory based on this intuition. According to him:

> The set of events relevant to determining the content of neural responses is the causally related set that has the highest statistical dependence with the neural states under all stimulus conditions (Eliasmith 2000, p. 34).

Eliasmith puts forward two conditions for a state R to represent S: R represents S iff there is a causal link and a statistical dependence relation between R and S.[2] But how does Eliasmith interpret the notion of 'statistical dependence'? He claims that "a statistical dependence between two events means that the occurrence of one event changes (either increasing or decreasing) the probability of the occurrence of the other event" (Eliasmith 2000, p. 69). There is a high statistical dependence between two states when the occurrence of one of them increases or decreases the probability of the other. Thus, in the context of cognitive science, Eliasmith holds that content is fixed by the positive statistical dependence of stimuli on responses and also by the statistical dependence of responses on stimuli (i.e. $P(R \mid S)$ and $P(S \mid R)$). That is what he calls taking the 'observer' and the 'animal's perspective', respectively (Eliasmith 2005b). A state represents[3] the entity with which it has a higher statistical dependency.

A precise summary and elegant formulation of this idea has also been defended by Usher (2001). Usher, who explicitly bases his approach on Shannon's notion of mutual information, proposes that R represents S iff (1) the mutual information that R carries about S is greater that the information R carries about any other entity and (2) the mutual information that S carries about R is greater than the information S carries about any other representation. More formally:

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. $P(S_i \mid R_i) > P(S_i \mid R_j)$[4]

[2]Note, however, that the element that is doing the real work in fixing content is the statistical dependence relation; the causal element is mainly introduced in order to avoid certain counterexamples, such as cases involving two mental states with a common cause (Eliasmith 2000, p. 59).

[3]A terminological note: in this essay, we call 'representational content' what Eliasmith calls 'referent' (he distinguishes 'referent' from 'content' understood in a different sense).

[4]The connection with Shannon's measure of information becomes clear once it is noted that 1 and 2 are simplifications of the following inequalities (Usher 2001, p. 321):

(a) $MI(R_i; S_i) = log \frac{P(R_i|S_i)}{P(R_i)} > log \frac{P(R_i|S_j)}{P(R_i)} = MI(R_i; S_j)$

(b) $MI(R_i; S_i) = log \frac{P(S_i|R_i)}{P(S_i)} > log \frac{P(S_i|R_j)}{P(S_i)} = MI(R_j; S_i)$

These two conditions are supposed to capture the two dimensions that are relevant for content determination: the backward and forward probabilities (corresponding to what Eliasmith calls 'observer' and 'animal perspectives'). The first condition claims that among all the entities that increase the probability of R_i occurring, S_i is the one that increases this probability most. So, for example, although the activation of certain neural population, which we can call 'DOG', can be triggered, under some circumstances, by dogs, wolfs, fat cats, etc., DOG represents dogs partly because among all the stimuli eliciting DOG, dogs are the entities that better predict tokens of DOG. In what follows, we will sometimes express this idea by saying that DOG 'tracks' dogs better. The second requirement does not compare stimuli but representational states. The idea is that R_i represents S_i only if R_i is the representational state that most increases the probability of S_i being the case. Here the probability that matters is the backward probability, i.e. conditionalized on representational states. For instance, suppose that there is a neural population that is sensible to all kinds of mammals (call it 'MAMMAL'). Even if MAMMAL is sometimes triggered by dogs, the second condition determines that DOG, but not MAMMAL, represents dogs, because it is more probable that there is a dog if DOG is instantiated than if MAMMAL is tokened: $P(Dog \mid DOG) > P(Dog \mid MAMMAL)$. Indeed, this inequality holds even if no other stimulus is as efficient in triggering MAMMAL as dogs are in the individual's environment.

Both Eliasmith's and Usher's proposals try to naturalize representations by appealing to high statistical dependency relations between representations and representata. A slightly different approach is suggested by Rupert (1999). Although he also analyzes representations in terms of probability relations between entities, he only considers forward probabilities (i.e. conditionalized on entities rather than on states) and he restricts his account to representations of natural kinds. On this account, representational content exclusively depends on whether members of S are more efficient in their causing R[5] than are members of any other kind. His account can be effectively considered a version of SGIT, in which condition 1 of SGIT is necessary and sufficient for a state to represent another state, once S_i and Sj are restricted to states involving natural kinds.

SGIT 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, since they reject Dretske's suggestion that the likelihood of the referent given the representation has to be one, these theories 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 is not in its extension (see below). Secondly, SGIT are also far more

As Usher notes (p.320), since the logarithm is a monotonic function, and we only make use of ordinal relations, we can rely on exp(MI) that provides the same expression but without the logarithm. As in 1 and 2 in INFO.

[5]Rupert does not formulate his approach in terms of representational *states,* but as applying to 'terms in a language of thought'. Nonetheless, since the exact nature of the entity that does the representing is irrelevant for our discussion, we describe all accounts as talking about states.

realistic than previous proposals in this tradition. Indeed, as they insistingly point out, these accounts might actually capture the way neuroscientists reason (Usher 2001, p. 320).[6] For instance, following Hubel and Wiesel's (1959) pioneering 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 might be extremely valuable in scientific projects (Eliasmith 2000, p. 71; Skyrms 2010a). And they also attempt to explain and vindicate the nature of representations attributed by cognitive scientists.

For these and other reasons, in recent years there has been a growing body of interest on informational theories (Skyrms 2010b; Birch 2014) and its relation to cognitive science (e.g. Piccinini and Scarantino 2010; Stegmann 2013, 2015). In what follows, however, we would like to argue that these theories fall prey to important difficulties. These objections suggest that the tools employed by SGIT are probably inadequate for analyzing representational states. If these arguments are on the right track, the prospects of SGIT would need to be seriously revised.

3 Problems for SGIT

We will present four objections against SGIT: they suffer from the problem of error, deliver proximal contents, do not allow for states with disjunctive contents or multiple states with the same content, and fail to provide a metasemantic account of representation. Our first goal is to argue that new informational theories face these difficulties. Secondly, we will show that the problem is rooted in some aspect of the notion of information. That result would strongly suggest that informational theories need to be supplemented with (or perhaps, replaced by) a theory with a different set of tools.

3.1 Error

Let us begin with the first difficulty of classical informational theories that SGIT were designed to solve: the problem of error. In this section we will argue that SGIT fail to fully address the problem of error that caused the dismissal of previous approaches in this tradition.

First of all, it is important to note that SGIT can indeed account for *some* error (so it provides a significant improvement over Dretske's approach). To accommodate some cases of misrepresentation, an account only needs to be compatible with the following two conditions: (1) R represents S and (2) $\neg S$. This is clearly possible on SGIT.

[6] The fact that INFO might capture the implicit assumptions made by neuroscientists in establishing hypotheses about representational relations does not mean that SGIT are purely descriptive theories. Eliasmith (2005b), for example, has criticized neuroscientists for excessively relying on what he calls 'the observer's perspective' (i.e. $P(R \mid S)$) and forgetting about the 'animal's perspective' (i.e. $P(S \mid R)$).

Indeed, any approach that requires a conditional probability below 1 automatically leaves room for some cases of error (Kreamer 2013). For instance, a theory according to which R represented S only if $P(R \mid S) \geq 0.8$, would certainly allow for some cases of misrepresentation.

However, whereas SGIT can actually account for occasional error and even for some forms of common misrepresentation (Usher 2001, p. 331, for instance, presents one such case), there are plausible cases of frequent misrepresentation that cannot be accommodated. More precisely, we will argue that they cannot account for some cases in which misrepresentation is more frequent that accurate representation.

Let us try to spell out this idea in detail and illustrate it with some examples.[7] Recall that, according to the first condition of INFO, R_i's representional content is partly determined by the state S_i that R_i tracks the best, i.e. for any other state S_j, $P(R_i \mid S_i) > P(R_i \mid S_j)$. To challenge this claim, we need to find some examples in which R_i represents S_i and, nontheless, $P(R_i \mid S_i) < P(R_i \mid S_j)$. Consider the case of mimicry; the dronefly *Eristalis tenax*, for example, is a Batesian mimic of the honeybee *Apis mellifera*: the former is a defenseless insect that has copied the appearance of honey bee (including its flight – Golding et al. 2001) to avoid being preyed upon. If the tokening of the predator's brain state R_i depends upon the appearance of the prey, and the dronefly mimics sufficiently well the appearance of the dangerous bee, we can assume that $P(R_i \mid bee) = P(R_i \mid dronefly)$. Indeed, this might lead to a case in which $P(R_i \mid bee) < P(R_i \mid dronefly)$ for the following reason: the fitness of the mimicked organism decreases when mimics expand (because when only few bee-looking animals are actually bees, it might pay predators to risk and prey them anyway – see Ceccarelli and Crozier, 2007); thus, bees could actually evolve in the direction of looking "less bee-like", and as a result $P(R_i \mid bee) < P(R_i \mid dronefly)$. This is a particular case of $P(R_i \mid S_i) < P(R_i \mid S_j)$, which should be troubling for INFO, since it seems that predators represent bees, although this representation tracks droneflies better.

Now, there is a similar scenario to shows that INFO is incompatible with some cases of systematic misrepresentation. Formally, systematic misrepresentation occurs when R represents S, $S \neq S^*$ and, nonetheless, $P(S \mid R_i) < P(S^* \mid R_i)$. In principle, this situation is fully compatible with the two conditions stated in INFO. Nevertheless, the problem pointed out in the previous paragraph suggests that there are some cases of misrepresentation that cannot be accommodated. For instance, suppose that both bees and droneflies cause the predator's brain state in roughly a similar proportion of cases, i.e. $P(R_i \mid bee) = P(R_i \mid dronefly)$. Let us say that $P(R_i \mid bee) = P(R_i \mid dronefly) = 0.6$. This is a forward conditional probability (i.e. conditionalized on states), whereas the problem of systematic misrepresentation involves backward conditional probabilities (i.e.conditionalized on representations). According to Bayes' theorem, to get that value we need to consider the marginal probabilities. Thus, suppose we know the marginal probabilities of bees and dronelies, and imagine that the latter are slightly more common than the former, e.g. $P(bee) = 0.4$ and $P(dronefly) = 0.6$. In that case, $P(bee \mid R_i) <$

[7] We would like to thank an anonymous reviewer for pressing us on this issue.

$P(dronefly \mid R_i)$.[8] In other words, this is a situation in which R_i intuitively represents bees, but most of the time it is tokened when there is a dronefly (and, obviously, $bee \neq dronefly$) so it is a systematic misrepresentation. However, since condition 1 is not satisfied, it is a case INFO cannot accommodate.

Indeed, as we see it, the difficulty is not just that in certain scenarios, such as the one depicted above, INFO delivers the wrong results. The problem is that representational content heavily depends on the marginal probabilities of states, and that seems to base content on the wrong kind of considerations. In general, R_i can only represent S_i consistent with false positives outnumbering true positives if $P(S_i) > \neg P(S_i)$. Basing content on comparisons between marginal probabilities seems inadequate; it as if you could change the content of the frog's mental state by simply killing some bees.

As a response, one might bite the bullet and defend that predators do represent the presence of droneflies. Rupert (1999, p. 336), for instance, at some point seems to suggest this strategy. He discusses a similar example in which females of certain species token a mental state to decide with whom to mate and claims that if this mental state has a higher correlation with a male of a different species, that might be what they in fact represent. Nonetheless, there are powerful reasons for resisting this move. In Rupert's example, the female's behavior (and, indeed the very existence of the representational mechanism) would be hard to explain unless it is assumed that it represents conspecific males. Similarly, in mimicry, unless R means *bee* (or *dangerous animal* or the like), the explanation of the organism's mental states and behavior would remain mysterious. Why do predators fail to prey on an insect when R_i is tokened if it means *dronefly* and these insects are harmless food? It seems that the only way to make sense of the whole process is by supposing that the brain state means something like *bee* (or the like).[9]

Alternatively, one could try to resist this objection by claiming that predators represent neither bees nor droneflies, but whatever appearance property they share (something like *bee-looking thing*). This reply, however, only might seem plausible if one thinks of predators without sophisticated cognitive capacities. but suppose that predators are human beings. Certainly we can think of bees, not only of bee-looking things (otherwise, this whole section would be unintelligible to you). However, it is entirely possible that most of the time we apply the concept BEE to droneflies and that around us droneflies are more common than bees. In any case, the suggestion that representational states actually track proximal states rather than distal ones points at a serious objection to INFO that will be discussed in the next section.

[8] According to Bayes' rule, $P(S_i \mid R_i) = \frac{P(R_i|S_i)(S_i)}{P(R_i)}$. Thus, to derive $P(bee \mid R_i)$ and $P(dronefly \mid R_i)$ we need to know $P(R_i)$. Fortunately, as Usher remarks (see foonote 4), in the present context this value is not required, because we are only interested in comparing $P(bee \mid R_i)$ and $P(dronefly \mid R_i)$, and in both cases the numerator is the same, namely $P(R_i)$. Thus, the fact that $P(R_i \mid bee)P(bee) < P(R_i \mid dronefly)P(dronefly)$ is enough for showing that $P(bee \mid R_i) < P(dronefly \mid R_i)$.

[9] As a reviewer suggested, at least one should grant the conceptual possibility of predators representing bees even if they frequently mistake droneflies for bees and the former outnumber the latter. Depending on one's metaphysical assumptions, the mere fact that this is conceptually possible might be enough for raising a problem for INFO.

Consequently, SGIT have problems accommodating some cases of frequent misrepresentation. Although occasional mistakes and some forms of regular misrepresentations are allowed, others are rendered impossible. Given that one of their primary motivations for SGIT was to fully account for error, this is a significant result. Furthermore, we argued that SGIT seem to base content attribution on the wrong kind of considerations. The root of the problem is not hard to identify: SGIT fail to leave room for this kind of mistakes due to their reliance on statistical dependencies. Thus, the problem derives from their central assumption.[10]

3.2 Distality

The second difficulty we will discuss is the *wrong distality attribution problem* and the best way to motivate it is to present a slightly different problem that SGIT do *not* have: the indeterminacy problem. A theory suffers from the indeterminacy problem when it underdetermines content attribution, that is, when there are just too many entities that could be represented according to it. Suppose, for instance, a naive informational account, according to which a mental state R represents S iff the presence of the state increases the probability of S. Since activity in our photoreceptor cells raises the probability of certain photons striking the retina, but also the probability of it being sunny, of the subject being awake, of shops being open and many others, the content of R would be highly indeterminate. So the indeterminacy problem seems to jeopardize this naive informational theory. In that respect, INFO fares much better than the naive approach, because it picks up the *single* state that has the greatest statistical dependence with the representational state. Unfortunately, this solution to the problem of indeterminacy has two striking unwelcome consequences: the wrong distality attribution problem, which will be developed here, and the problem of ambiguous representations, which will be presented in the next section.

Consider the Fusiform Face Area (FFA), which is usually thought to represent faces (Kanwisher et al. 1997; Desimone 1991). Suppose we discover that a certain neural network R in the FFA selectively fires with significant intensity when there is a face and also that, given that R is active, the entity that is more likely to be present is a face. One might think these observations suffice for establishing the fact that the brain state represents *face* according to SGIT. Unfortunately, it is unclear that SGIT can deliver this result. Consider, for instance, the set of neuronal structures in the thalamus that are active whenever there is a face in front of the subject. If R has the highest statistical dependence with faces, it will also normally have the highest statistical dependence with these neuronal states in early vision. Thus, SGIT would entail that this activity in FFA represents neuronal activation in another part of the brain. This is of course an extremely counterintuitive result. Indeed, even if there was some principled way of excluding other brain states from being represented, other inadequate contents such as *face-looking thing* could probably not be avoided.

[10]It has been argued that teleological theories are also incompatible with some forms of systematic misrepresentation (Mendelovici 2013, 2016). For a response, see Artiga (2013).

More generally, a theory suffers from the wrong distality attribution problem when it systematically delivers content at the wrong level of distality.[11] Since mental states will normally have a higher statistical dependence with most proximal stimuli, SGIT would tend to deliver contents that are too proximal. This result is at odds with our intuitions and seems to be in tension with the claims made by cognitive scientists (see Sebastián and Artiga forthcoming).

Of course, supporters of SGIT can insist that, if a representational state actually means *face,* it will surely have higher statistical dependence with faces than with face-looking things. However, this is a doubtful assumption. First, for the reasons provided in the previous section, it is not obvious that a representation of faces needs to correlate better with faces than with non-faces. Secondly, since we usually identify faces by detecting face-looking things, it is not unreasonable to suppose that the correlation with proximal stimuli is at least as good as the correlation with distal stimuli. Similarly, given that we identify face-looking things by detecting certain features, the correlation with the latter will be stronger. And so on. This reasoning can be iterated until the most proximal state that actually triggers the brain state (which in many cases will be another mental state) holds. Consequently, SGIT clearly suffer from the wrong distality attribution problem.

Finally, one might think that the wrong distality attribution problem is less troubling for Rupert's account, since he restricts his approach to representations of natural kinds. This is doubtful. His proposal still faces a daunting dilemma, whose horns depend on whether the set of candidate properties that give rise to the wrong distality attribution problem qualify as natural kinds or not. If, for instance, light photons or certain kind of neuronal states in early vision count as natural kinds, the problem remains in all its force.[12] If they do not qualify as natural kinds, then the approach clearly excludes too much: these entities are sometimes represented by cognitive states, so ruling them out *by definition* is clearly inadequate. A different answer to this problem is briefly sketched by Eliasmith. He suggests an additional requirement: 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 (Eliasmith 2005a, p. 1047; Eliasmith 2000, p. 59-60). However, this proposal is still unsatisfactory. First of all, many think that computations are defined over representations. For this reason, 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

[11] Some employ the label 'distality problem' to refer to this difficulty, but this expression is also frequently used for version of the indeterminacy problem (e.g. Neander 2017, ch.9; Schulte 2018). To avoid any sort of misunderstanding, we call this objection the 'wrong distality attribution problem'.

As a reviewer pointed out, one might question whether in all cases the problem we present concerns more distal vs. less distal features (consider, for instance, the contrast between faces and face-looking things). We adopted this terminology here because it is very usual in the literature and in many cases there is clearly a contrast in distality (e.g. faces vs. neuronal patterns). Nonetheless, the name is not important; the key point is that the theory delivers the wrong content, but not in virtue of it being indeterminate.

[12] Actually Rupert (1999, p. 340) accepts a extremely liberal approach to natural kinds, according to which "natural kinds are any kinds that successful non-intentional science finds theoretically interesting and useful". Thus, he probably faces the first horn of the dilemma.

this is precisely what this condition is supposed to establish. Secondly, this approach is also too exclusive and too inclusive at the same time. One the hand, it still includes too much because the problem is not only caused by neuronal states, but also by external entities (e.g. photons, object-looking things, etc...). On the other, it excludes too much because some neuronal states indeed represent other states of the brain or the body (Damasio 2010; Rosenthal 2005; Prinz 2004). For a more detailed discussion, see Sebastián and Artiga (forthcoming).

How could one attempt to solve this problem? We think progress could be made by appealing to some other notions such as function. In a nutshell, the idea is that although a given mental state M has a higher statistical correlation with a proximal feature F, it might still be its function to represent something more distal. Taking this path, however, would have significant consequences: if mental states have a higher statistical dependence with their most proximal cause while their function and content concerns a distal feature, it is not obvious that the notion of information is actually playing any important role in a theory of representation. In other words, if the mental states tend to have a stronger correlation with their proximal causes, but tend to have distal contents, then it is unclear whether the former can be used to provide an account of the latter. Of course, much more should be said to make this thought fully compelling; one would need to spell out in detail this notion of function and show how it can deliver the right level of distality. Nonetheless, we think that supporters of SGIT should seriously consider the possibility that the wrong distality attribution problem might show that the notion of higher statistical dependence is actually an inappropriate tool for a theory of content.

3.3 Multiple Representations and Contents

We saw that the fact that SGIT restrict representational content to the *single* entity that has a higher statistical dependence with a mental state helps them reduce the indeterminacy faced by the naive informational approach, but it gives rise to the wrong distality attribution problem. In this section we would like to highlight a second negative consequence. It is a platitude that, in many cases, the relation between representations and their contents is not one-to-one, but many-to-one or one-to-many. As a matter of fact some representations have multiple referents and the same referent is sometimes represented by different states. However, SGIT render these facts impossible by definition. By requiring content to be determined by the single entity that has a higher statistical dependence, SGIT have difficulties in accounting for representations with multiple contents and contents shared by multiple representations. This problem derives from each of the two conditions included in INFO.

An example from our visual system might help illustrate the problem. Consider the case of *metamerism*. The human eye contains only three types of cone cells, which are responsible for color vision. Each of these these types of cell respond to the cumulative energy from a broad range of wavelengths. Now, as it happens, different combinations of light across all wavelengths can produce an equivalent receptor response. Consequently,—on the assumption that colors are surface reflectances—there are colors (called 'metamers') that despite having different spectral power

distributions produce the very same neural network activation in the visual cortex. Call 'C_1' and 'C_2' two such metamers. Consider the neural network, R_i, in the visual cortex that satisfies condition 2 of INFO: $P(C_1 \mid R_i) > P(C_1 \mid R_j)$ for every $R_j \neq R_i$. Thus, R_i is the network that most strongly increments the probability of C_1. Imagine also that $P(R_i \mid C_1) > P(R_i \mid C_2)$. Condition 1 is also satisfied, and so, according to the theory, R_i represents C_1. However, intuitively, the content of R_i is rather $C_1 \vee C_2$. The theory cannot provide this result, at least as there is a slight difference in probabilities between C_1 and C_2. Under these conditions, organisms cannot represent C_2 , even if C_2 were in fact the most common color after C_1 in the environment.

If SGIT face difficulties with representations with a disjunctive content, an analogous problem arises in the context of multiple representations with the same content. In particular, according to condition 2 of INFO, R represents a stimulus S only if there is no other state R_j, such that $P(S_i \mid R_j) > P(S_i \mid R_i)$. That means that only one representation can have S as its content. Thus, according to these theories by definition two representations cannot have the same referent, yet we do have different representations with a common reference. A striking example is provided by redundant processing. Redundancy is the duplication of critical functions with the purpose of increasing the reliability of the system. For example, if vision is lost in one eye we do not become blind (although depth perception is impaired) and often the same information is presented to both eyes. Redundancy is an important field of study in cognitive science. There is strong evidence, for instance, that redundant presentation of information across modalities recruits attention and enhances learning (what is called the 'intersensory redundancy hypothesis' [IRH]: Bahrick et al. 2002; Bahrick and Lickliter 2000; Bahrick et al. 2004). As Bremer and colleagues (2012, p. 955) suggest:

> According to the intersensory redundancy hypothesis, optimum conditions for deriving benefit from provision of multisensory information would be those in which both visual and auditory information provide congruent information about an object's trajectory. Under such conditions, visual and auditory information would specify the object's trajectory redundantly, and so could be expected to enhance perception of trajectory continuity as the object passed behind an occluder.

Whatever the merits are of the Intersensory Redundancy Hypothesis, it is a coherent theory whose truth or falsity should be assessed empirically. However, SGIT seem to be incompatible with its truth. To take a particular example, consider two different networks representing the object's trajectory, one in the visual cortex, V_t, and one in the auditory cortex, A_t. According to the Intersensory Redundancy Hypothesis, this redundancy enhances perception of the trajectory when the object passes behind an occluder. In certain cases of maximal congruence, V_t and A_t provide the same information and both represent the same trajectory, T. The problem is now straightforward: if, on the one hand, $P(T \mid V_t) > P(T \mid A_t)$, then A_t does not represent T, and if, on the other hand $P(T \mid V_t) < P(T \mid A_t)$, then V_t does not represent T. What SGIT cannot accept is that both neural networks represent the trajectory of the object.

Anticipating a similar problem, Rupert (1999, p. 349-350) suggested the following amendment in the theory:

> R has a truly disjunctive extension when the following conditions, (a), (b), and (c), are met: (a) two or more natural kinds have equal or roughly equal success rates relative to t; (b) no other natural kind has a success rate substantially higher than those kinds whose success rates are equal or roughly equal relative to t; and (c) the gap between the group of success rates at the top and those farther down is substantial.

Generalizing this idea, the suggestion is to qualify condition 1 of INFO and accept that the required probability need not be strictly higher; it suffices if $P(R_i \mid S_i) \geq P(R_i \mid S_j)$. In this way, SGIT could account for some ambiguous representations. A similar modification of condition 2 (i.e. $P(S_i \mid R) \geq P(S_i \mid R_j)$) could leave room for multiple representations with the same content.

Does this suggestion provide a convincing way out to the problem? It is unclear that this additional clause can give the right results. For one thing, there are surely many cases of ambiguous or redundant representations in which the probabilities of different stimuli or representations are unequal. For another, equal probabilities does not ensure coreference, precisely because of the existence of regular misrepresentations. In Rupert's formulation, the efficiency of the entities that relate to different meanings in an ambiguous representation has to be equal or roughly equal, i.e. R refers to two different entities iff a subject regularly and consistently applies R to S_i and S_j and this fact contrasts with how often S_i and S_j cause other representations. Yet this suggestion does not take the possibility of recurrent misrepresentation seriously enough. In particular, there is nothing in this approach that could distinguish ambiguous concepts from systematic mistakes. Taking the example discussed earlier, consider the predator's representations of bees: the predator might systematically confuse droneflies for bees and that does not mean that its representation is ambiguous; it just means that it is wrong roughly half of the time. Unfortunately, an exclusive reliance on statistical dependencies renders it unable to make this distinction. Thus trying to leave room for multiple contents and redundancy by simply relaxing the conditions for representing, does not ensure that the right disjunctive contents are predicted.

As a result, SGIT are probably unable to account for genuine cases of ambiguity and redundancy. Again, other notions like functional role or teleofunction (Millikan 2000) might contribute to solving this problem. But statistical dependence does not seem to be the right tool for the task at hand.

3.4 Metasemantics

The last problem of SGIT is that they do not even address the most pressing question for a theory of representation, namely, what makes certain states representational. Let us elaborate.

Generally, the discussion on naturalistic theories of content has failed to make explicit an important distinction between two different goals. First of all, a theory of representation must explain why a state R represents S_1 rather than a different state

S_2. For instance, one might attempt to explain why certain neuronal activity in the striate cortex indicates the presence of a vertical line rather than a circle. Secondly, a satisfactory theory also has to provide an account of why R represents something at all. According to this second way of addressing the problem, the goal is not to explain why neuronal activity is supposed to detect vertical lines rather than something else, but why this state is a representation at all. Let us call the first kind of theory 'semantic' and the second one 'metasemantic' (Artiga 2016).

Now, that distinction is important because solving the problem of intentionality requires providing a semantic *and* a metasemantic theory. If we want a theory to fully explain why certain states are representational and how their content is determined, we need to provide both kinds of theories. Merely explaining why a certain state represents lines rather circles does not completely dispel the mystery posed by representational phenomena, since there is still a fundamental question that remains unanswered: why does this state represent something at all? Unless this issue is addressed, we will lack a solution the problem of intentionality. The following comparison might be illuminating: the problem of consciousness cannot be completely solved by merely providing a theory of why a subject experiences blue rather than red (e.g. by mentioning the fact that different parts of the brain are active or that they represent a different content). On top of that and more importantly (Kriegel 2009), we need to answer the question of what makes it a conscious state. We abstract from the particular ways having different experiences feel and concentrate on the problem of what makes it the case that having a conscious experience feels any way at all. Likewise, the problem of intentionality concerns the very nature of representational states. Providing a semantic theory is of course a step in the right direction, but it leaves unresolved a major question.

With this distinction in mind, it should be obvious why informational theories do not provide a metasemantic theory of representation. Even if they correctly identified the conditions that should be satisfied for a representational state R to refer to S_1 rather than to S_2, they do not put forward any criterion for determining when a certain state is indeed a representation. Informational theories offer semantic theories, but fail to deliver metasemantic ones.

Of course, it could be replied that the fact that there is an aspect of the project that these theories have not yet addressed does not mean that they could not do it. Future work might fill this gap without abandoning a purely informational framework. Unfortunately, the problem seems to go much deeper than that: not only have informational theories actually failed to provide a metasemantic account, but the informational tools they employ seem to be inadequate for carrying it out. The reason is well-known: informational relations are ubiquitous. Any state that is subject to be in a causal chain—that is, any state that is not outside the causal order—carries information about its possible causes and effects (among other things). However, we do not want to maintain that representations are ubiquitous. Thus, the concept of statistical dependence is not fine-grained enough for distinguishing what is a representation from what is not. *Carrying information*, like the properties *having a certain weight* or *being an object* are just too coarse-grained for the task at hand. What at first glance might look like an oversight actually constitutes an important gap in the theory.

Certainly, we do not want to deny that there are metasemantic theories compatible with informational accounts. Again, one could add functional roles, teleofunctions or other tools in order to define what representations actually are. But, in any case, these hybrid theories would provide a solution to the metasemantic problem only if they go beyond the notion of information. Consequently, purely informational theories (such as SGIT) are unlikely to provide a fully convincing account of representational phenomena.

4 Conclusion

Summing up, in this paper we have shown that Scientifically Guided Informational Theories (SGIT) face four important problems. Furthermore, we have argued that the reasons they fail to overcome these difficulties are deep: because of their exclusive reliance on information, they simply lack the resources for providing satisfactory solutions. Thus, new-wave informational theories are unlikely to succeed in the project of providing a theory of representation in the context of cognitive science.

Nonetheless, it is worth stressing that the arguments suggested here are not intended to show that the notion of information is useless. Cognitive scientists heavily rely on informational measures and the intuition that this notion captures something important about cognition is a powerful one. Our arguments are not intended to suggest that the notion of information should be eliminated, but rather that the connection between information and representation needs to be reassessed. In particular, attributions of representations in scientific practices might not just rely on statistical dependence (at least not always, as some of the examples we have presented show). Furthermore, other notions such as 'function' might need to be added to address the previous worries. Consequently, additional considerations have to be made explicit and placed on the table for a proper assessment. Whether this should lead to a revision of our current scientific methodology is an important question we leave for future research.

Acknowledgements We would like to thank Axel Barceló, the NCH Mind and Brain conference 2016 and two anonymous referees for their helpful comments and criticisms. Financial support was provided by a Postdoctoral Fellowship at the MCMP-LMU, the fellowship 'formación postdoctoral' from the Ministerio de Economia y Competividad, the UNAM-DGAPA-PAPIIT programs

References

Artiga, M. 2013. Reliable misrepresentation and teleosemanics. *Disputatio*. 37.

Artiga, M. 2016. Liberal representationalism. A deflationist defense. *Dialectica.* 70(3): 407–430.

Bahrick, L., and R Lickliter. 2000. Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology* 36: 153–186.

Bahrick, L., R. Flom, and R. Lickliter. 2002. Intersensory redundancy facilitates discrimination of tempo in 30 month old infants. *Developmental Psychobiology* 41: 352–363.

Bahrick, L., R. Lickliter, and R. Flom. 2004. Intersensory redundancy guides infants selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science* 13: 99–102.

Birch, J. 2014. Propositional content in signalling systems. *Philosophical Studies* 171-3: 493–512.
Bremner, J.G., A. Slater, S.P. Johnson, U.C. Mason, and J. Spring. 2012. The effects of auditory information on 4 month old infants perception of trajectory continuity. *Children development* 83(3): 954–964.
Chun, M., N. Kanwisher, and J. McDermott. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience* 17(11): 4302–4311.
Damasio, A. 2010. *Self Comes to Mind: Constructing the Conscious Brain*, 1st edn. New York: Pantheon.
Desimone, R. 1991. Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience* 3: 1–8.
Dretske, F. 1981a. *Knowledge and the Flow of Information*. Cambridge: The MIT Press.
Dretske, F. 1981b. *Knowledge and the Flow of Information*. Cambridge: MIT Press.
Eliasmith, C. 2000. How neurons mean: a neurocomputational theory of representational content. Unpublished Dissertation, Washington University in St. Louis.
Eliasmith, C. 2005a. Neurosemantics and Categories. In *Handbook of Categorization in Cognitive Science*, eds. H. Cohen, and C. Lafebvre. Amsterdam: Elsevier.
Eliasmith, C. 2005b. A new perspective on representational problemss. *Journal of Cognitive Science* 6: 97–123.
Eliasmith, C. 2013. *How to build a brain: A neural architecture for biological cognition*. New YOrk: Oxford University Press.
Godfrey-Smith, P. 1991. Signal, detection, action. *Journal of Philosophy* 88(12): 709–722.
Hubel, D.H., and T.N. Wiesel. 1959. Receptive fields of single neurones in the cat striate cortex. *Journal of Physiology* 148: 574–59I.
Kraemer, D. 2013. Against "soft" statistical information. *Philosophical Psychology* 28(1): 139–147.
Kriegel, U. 2009. *Subjective Consciousness: A Self-Representational Theory*. New York: Oxford University Press.
LeDoux, J. 2003. The emotional brain, fear, and the amygdala. *Cellular and Mollecular Neurobiology* 23(1): 727–738.
Mendelovici, A. 2013. Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies* 165(2): 421–443.
Mendelovici, A. 2016. Why tracking theories should allow for clean cases of reliable misrepresentation. *Disputatio* 8(42): 57–92.
Millikan, R.G. 1989. Biosemantics. *The Journal of Philosophy* 86: 281–297.
Millikan, R.G. 2000. On Clear and Confused Ideas. Cambridge University Press.
Neander, K. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge: MIT Press.
Oehman, A., and S. Mineka. 2001. Fears, phobias and preparedness: Toward an evolved module of fear and fear learning. *Current Biology* 17(13): 129–33.
Prinz, J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University 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.
Scarantino, A., and G Piccinini. 2010. Information processing, computation, and cognition. Journal of Biological Physics.
Schulte, P. 2018. Perceiving the world outside: How to solve the distality problem for informational teleosemantics. *Philosophical Quarterly* 68(271): 349–369.
Sebastián, M. Á., and M. Artiga. forthcoming. Can informational theories account for metarepresentation?. *Topoi*.
Skyrms, B. 2010a. *Signals: Evolution, Learning, & Information*. Oxford: Oxford University Press.
Skyrms, B. 2010b. *Signals: Evolution, learning and information*. Oxford: Oxford University Press.
Stegmann, U. 2013. Animal Communication Theory. Information and Influence In ed. U. Stegmann, *A primer on information and influence in animal communication*. New York: Oxford University Press.
Stegmann, U. 2015. Prospects for probabilistic theories of natural information. *Erkenntnis* 80: 869–893.
Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language* 16(3): 331–334.