Nomic Locking:
A New Informational Theory of Content

Verónica Gómez Sánchez
March, 2020

(Early draft, please do not cite)

Introduction

According to the representational theory of the mind (RTM), propositional attitudes like beliefs and desires are realized by computational relations between subjects and structured mental representations. These mental representations have causal powers as well as semantic contents: they represent the world as being a certain way, and they can be true or false depending on whether the world is that way or not.

One central question for RTM concerns the metaphysical basis for semantic contents: in virtue of what does a mental structure come to stand for a mind-independent aspect of reality—be it a thing, a property or a proposition? Many philosophers and scientists have thought that the answer to this question has to do with the existence of certain robust correlations connecting mental representations and their worldly correlates. The idea, in broad strokes, is that representations get their meanings by being deployed in accordance with the presence and absence of certain features in the environment. A representation gets to be about an object, property or proposition in virtue of standing in a kind of ‘tracking’ relation to it, where ‘tracking’ is to be defined in terms of scientifically respectable notions like correlation, causation, and/or probabilistic dependence.

Despite the intuitive appeal of the general idea, attempts to accomplish the promised reduction of ‘tracking’ to scientifically respectable notions face seemingly insurmountable problems. Traditional informational approaches have trouble accounting for a basic feature of human cognition: misrepresentation. In addition, they struggle to yield sufficiently determinate contents. The tracking relations they invoke connect mental representations to more entities than semantic relations intuitively do.

In light of these difficulties, many philosophers now believe that tracking accounts of content cannot succeed unless they are supplemented with a teleological component\(^1\). The idea, roughly, is that we

\(^1\) See Dretske (1986, 1988)
need to distinguish between the worldly entities/states that our representations track, and those that our representations are supposed to track (or have the function to track). Only the latter can be genuine semantic contents.

This paper develops an alternative solution to the two problems stated above, which does not appeal to natural teleology. I propose that interpretation functions are devices for concisely summarizing patterns of informational connections between elements in a representational system and environmental states of affairs. I then argue that, given plausible theoretical constraints on good summaries, we can explain how misrepresentation is possible and how content determinacy is achieved.

1. Informational semantics and its discontents

In Knowledge and the Flow of Information (1981), Fred Dretske introduced the idea that the semantic contents of beliefs may be understood in terms of a purely probabilistic notion of information. In this section, I will present this account and two well-known problems with it, which I take to be the main reasons for its unpopularity: the problem of misrepresentation and the problem of semantic indeterminacy. The goals of this discussion will be to (i) highlight the features of Dretske’s original proposal which gave rise to these problems, (ii) identify some alternative routes that Dretske could have taken. This will motivate the improved version of informational semantics that I develop in section 2.

In general, we can think of informational accounts as consisting of two elements. First, they impose a necessary tracking condition on representation. This condition states that, if a mental representation R represents \( p \), that there is a relation of causation and/or robust covariation obtaining between R and \( p \). Second, alongside the tracking condition, a ‘selecting condition’ is needed: a condition which selects among all the propositions that R ‘tracks’ a single proposition that is its semantic content.\(^2\)

In Dretske’s account, the necessary tracking condition is stated in terms of the probabilistic notion of ‘semantic information’, and the ‘selecting condition’ invokes the notion of specificity. The complete account has the following form:

Representational vehicle \( r \) has semantic content \( p = \)

(i) \( r \) carries the information that \( p \).
(ii) \( p \) is the most specific piece of information that \( r \) carries.

\(^2\) Or a few propositions in cases where the representation is vague. To simplify the discussion, I start by supposing that there is no vagueness: the contents of all representations are perfectly precise. It should be clear, however, that none of the arguments I give crucially depend on this assumption.
Dretske cashes out the notion of information in probabilistic terms: \( r \) carries the information that \( p \) if and only if conditional probability of \( p \) given that \( r \) is tokened is maximal—i.e., \( P(p \mid r \text{ is tokened}) = 1 \).

The notion of probability that features in this definition is an objective notion (note that, if the probabilities in question were subjective credences, the account would not be genuinely reductive). Dretske suggests that these objective probabilistic connections are underwritten by ‘nomical regularities’. More specifically, if \( r \) carries the information that \( p \), then there must be (i) properties/types \( R \) and \( P \) which \( r \) and \( p \) fall under (respectively), (ii) and a nomical regularity which ensures that every instance of \( R \) is accompanied by a (suitably related) instance of \( P \).

Let’s apply condition (i) to a simple example. Consider a belief of the form \( X \text{ IS GREEN} \), which a subject \( S \) forms during a direct perceptual encounter with a particular object \( x \). According to Dretske’s proposal, \( X \text{ IS GREEN} \) has the proposition \( x \text{ is green} \) as its semantic content only if the objective probability that this proposition holds conditional on \( S \) having the belief \( X \text{ IS GREEN} \) must equal 1. This, in turn, requires that there be a lawlike regularity ensuring that, if \( S \) has the belief \( X \text{ IS GREEN} \), then \( x \) is in fact green.

There’s an obvious problem with this condition. If there is a lawlike regularity ensuring that \( x \) is green whenever \( S \) believes \( X \text{ IS GREEN} \), then \( X \text{ IS GREEN} \) can never falsely represent \( x \) as being green. But this can’t be right. Clearly, it is nomologically possible for \( S \) to misperceive \( x \)’s color: she might be hallucinating, unknowingly wearing tinted glasses, under abnormal lighting conditions, etc. This point generalizes: it is possible for \( S \) to mistakenly hold any belief that she is capable of entertaining, but condition (i) as stated cannot accommodate this.

To make room for misrepresentation Dretske introduces a crucial modification to the account. He suggests that the probability distribution with respect to which information is defined must incorporate some nomologically contingent information about ‘channel conditions’. In the above example, these channel conditions would include the information that the lighting is normal, the subject is not wearing color-altering glasses or running a high-fever, etc.

---

3 While Dretske formulates his account in terms of event types and tokens, I’m formulating the account in terms of propositions to make the continuity with my positive proposal apparent. Nothing important turns on this choice, however. Dretske was clearly working with a fine-grained conception of events, and if there are any metaphysical differences between fine-grained events and propositions, they won’t matter for our purposes.

4 Strictly speaking, a false proposition can have probability 1 (e.g. the proposition that a continuous random variable does not take the value it has actually has). But we can safely ignore such cases, since they won’t save Dretske’s account.
The resulting notion of information is then:

\[ r \text{ carries the information that } p \text{ iff for some } \text{‘channel condition’ } CC \text{ that actually co-occurs with } r, \]
\[ P(p \mid r \text{ is tokened } \& CC) = 1. \]

According to Dretske, these channel conditions are the conditions which obtain during the ‘learning period’: the time during which a mental representation acquires its content. Once a given representation has acquired the meaning that \( p \)—by being repeatedly tokened in circumstances where it carries the information that \( p \)—it can then be mistakenly tokened under new circumstances. For example, \( X \text{ IS GREEN} \) can be mistakenly tokened under the influence of hallucinogenic drugs after the learning period has ended.

As Loewer (1987) persuasively argued, this appeal to ‘channel conditions’ and/or a ‘learning period’ threatens to undermine Dretske’s reductive ambition. If by ‘channel conditions’ Dretske means whatever conditions are in place during the learning period, what marks the end of the learning period? And what other principled definition of ‘channel conditions’ can be given, other than ‘the conditions under which the system to tokens concepts correctly’? The concern here is that our grip on both of these notions is dependent on our pre-theoretical understanding of what a system is supposed to represent.

The standard strategy for dealing with this problem (which Dretske pursues in later work (1986, 1988)) is to replace the notion of channel conditions with a notion of ‘conditions for normal functioning’, where the latter are understood in terms of a process of design, selection or learning that explains the existence of the relevant representational capacities. Pursuing this strategy leads to a version of teleosemantics (Millikan (1984, 1989), Papineau (1984, 1993), Neander (2017)) But there is a second possible strategy that has received less attention: eliminate the need for a notion of channel conditions altogether, by making the tracking condition less demanding; in other words, drop the requirement that a representation be capable of perfectly tracking its content under some specified (channel/normal) conditions.

I think there are at least two reasons why this alternative route has been neglected. Firstly, the most straightforward ways of weakening the tracking condition are not very promising. We could say, for instance, that \( r \) carries the information that \( p \) if and only if \( P(p \mid r) \) is greater than some threshold \( t \), where \( t < 1 \). But any threshold we choose will be implausibly arbitrary.\(^5\) And, moreover, there doesn’t seem to be any minimum bound on how objectively unlikely a proposition can be conditional on our believing it. (Anecdotally: I used to believe that, whenever one set is a proper subset of another, the first has fewer members than the second.)

\(^5\) Millikan (1989) raises a similar arbitrariness worry for statistical approaches to the notion of ‘normal conditions’. 
Secondly, any weakening of the tracking condition creates more work for the selecting condition to do. The easier we make it for tracking relations to obtain, the more candidate meanings our selection condition must rule out. My goal in §2 is to show that we can make do with a very weak tracking condition that avoids the above problems, by coupling it with a much more powerful selection condition than Dretske’s. This strengthened selection condition will also solve a second problem for Dretske’s account—‘the problem of indeterminacy’—which I now turn to.

A central challenge for tracking approaches is that intuitively univocal representations track too many different propositions. The second clause in Dretske’s account is designed to compensate for this. But, as we will see now, it fails to accomplish its aim.

Suppose that, after taking a proper look of some object (under normal lighting conditions), you form the belief that this object is green. There is then a mental representation of the form X IS GREEN playing the role of belief in your cognitive system. By Dretske’s lights, the fact that this belief-representation has the content that x is green has to do with the fact that it carries that information. But, clearly, there is a lot of other information that your representation carries: the information that x is colored, that x reflects light, that x is green-or-blue, that x’s surface area is non-zero, etc. What privileges the proposition X IS GREEN over all of these nearby alternatives? I will call this the ‘selection’ problem for informational accounts of content: how, among all the features of reality ‘tracked’ by a representation, is one selected as its content?

Dretske intended condition (ii) to solve the selection problem. X IS GREEN is the content of X IS GREEN because it is the most specific piece of information that the representation carries (in Dretske’s terminology, the information that the representation carries in ‘digital form’). The notion of specificity that Dretske has in mind may be characterized in terms of metaphysical entailment: p is more specific than q if and only if necessarily, q obtains if p does, but q could obtain without p. That is: p necessitates q but not vice-versa. Since the proposition that x is green necessitates but isn’t necessitated by the propositions X IS colored, X reflects light, X IS green-or-blue, and X’S surface area is non-zero, the account predicts that none of these associated propositions gets to count as the representation’s content.

This specificity condition is not, however, a general solution to the selection problem, for not every piece of information carried by a representation (in Dretske’s sense) is entailed by what we intuitively consider to be its content. Two types of cases have been used to illustrate this problem: proximal stimuli and gerrymandered variants.

Consider, first, proximal causes. My ability to perpetually detect the color green depends on the fact that green objects in my visual field have a characteristic effect on my visual system: a pattern of activation of my photoreceptors, say. Call this pattern G. Consider a perceptual judgment: THIS IS GREEN where THIS demonstratively picks out some object x in my visual field. This belief will not
only carry the information that $x$ is green, but also the information that, a few milliseconds before the perceptual judgement was made, my photoreceptors exhibited the activation pattern $G$. This information is clearly not necessitated by the content of the belief: it is metaphysically possible for the object in question to be green without my photoreceptors registering this fact. But Dretske’s condition (ii) requires that any information carried by a belief be entailed by the content of that belief. So the content of my belief THIS IS GREEN cannot be the proposition that $x$ is green. Instead, the content will have to be a conjunction along these lines:

$$x \text{ is green and my photoreceptors exhibit activation pattern } G \text{ and …}$$

And this is clearly the wrong prediction. Thus, condition (ii) is not an adequate solution to the selection problem.

In response to this kind of concern, Dretske emphasizes that our beliefs typically have many different possible proximal causes, which means that a single belief won’t carry a particular proximal cause as its content. Suppose, for example, that I’m wearing color-inverting glasses. I may still form the belief THIS IS GREEN without the corresponding activity in my photoreceptors, but, rather, on the basis of my knowledge about how the glasses work (or on the basis of testimony). Dretske may say that this shows that the belief does not really carry information about the pattern of activation of my photoreceptors: the probability that my photoreceptors instantiate $G$ is, in fact, not maximal given that I have the belief. But this response is unsatisfactory. Firstly, the imagined case with the funny glasses seems exactly the kind of case where Dretske should want to say that channel conditions don’t obtain. Secondly, as Loewer (1987) points out, even if the belief does not carry the information about photoreceptors, it surely carries some disjunctive information about proximal causes: that the belief was either caused by such and such property of photoreceptors, or such and such auditory stimulus, or…

Besides proximal stimuli, ‘gerrymandered’ properties also raise a version of the selection problem. This has been discussed widely in connection to Quine’s (1960) arguments for semantic indeterminacy. Take the proposition: ‘There’s a rabbit there’. If this belief carries the information that there’s a rabbit at a given egocentric location, it also carries the information that there’s an undetached rabbit part at that location. Since these two propositions are true at the same possible worlds, neither is more specific than the other. Condition (ii) will be unable to privilege the intuitively correct content.6

On very coarse-grained conceptions of propositions, this result may be tolerable. But there are other gerrymandered contents which are not so easily embraced. Consider the following variant of Goodman’s predicate ‘grue’ (Gomez Sanchez, ms):

\[ \text{footnote}{6} \text{ See Gates (1996) for discussion of Quinean indeterminacy in the context of informational semantics.} \]
An object is grue* if and only if it is in the field of vision of some sentient being at some point in time, or blue and nowhere near sentient beings at any time. If the belief X IS GREEN carries the information that x is green, it also carries the information that it is grue*. It follows from the definition of grue* that whenever a sentient being encounters a green object, that object is also grue*. This means that the objective probability that any perceptual system comes into contact with a green object that is not also grue* is 0. Therefore, X IS GREEN carries the information that the x is grue*. But x’s being grue* is not necessitated by its being green.

The moral of this discussion is that, while Dretske’s selecting condition helps exclude some candidate contents, it does not rule out proximal causes or gerrymandered alternatives. A satisfactory solution to the selection problem requires extra resources.

2. The account

In the rest of the paper, I introduce a new version of informational semantics, and argue that it fares much better than Dretske’s view: it avoids both the problem of misrepresentation and the problem of indeterminacy, and does so without appeal to a notion of channel conditions. This account will develop the following general idea:

Interpretation functions are devices for concisely summarizing the informational connections between elements in a representational system and environmental states of affairs.

The overall picture is as follows: informational connections between mental representations and propositions in the world are abundant. While these connections are the ultimate basis for representation facts, the way in which they determine representation facts is less direct than we might initially hope. To get representation facts, we need to construct a certain kind of ‘summary’ of all the informational facts. The best summary involves a mapping from mental representations to worldly states, and it is this mapping that gives the contents of the mental representations.

2.1 Informational Links

The notion of information that I’ll be working with is importantly different from Dretske’s. As I said before, I think it is a mistake to narrow down the candidate contents to those propositions that are maximally likely given a representational state. I will retain Dretske’s connection between conditional probability and information, but I will work with a degreed notion of information.

I take the strength of the informational connection between a mental sentence-type R and a content ρ (for subject S at time t) to depend on two probabilistic relations that come in degrees:
(i) The conditional probability of $p_t$ given that $S$ tokens $R$ at $t$.

(ii) The degree to which $S$ tokening $R$ at $t$ evidentially supports $p_t$ (i.e., raises the probability of $p$).

A complete account must settle the question of how these two components are to be balanced to determine the overall strength of an informational connection. Here I leave this as an open question, to be set by future investigation into the nature of representation.

I think that, together, these two components do a good job at capturing the intuitive notion of indication or ‘natural meaning’. For example, a barometer reading indicates that the atmospheric pressure is such and such because (i) the atmospheric pressure is very likely to be as the barometer indicates, and also because (ii) the barometer reading provides strong evidence for the corresponding proposition about atmospheric pressure. Similarly, the presence of smoke in a forest indicates a fire because (i) the conditional probability of a fire given the presence of smoke is high, and because (ii) the presence of smoke is strong evidence for the presence of a fire.

For concreteness, I will be appealing to a standard probabilistic notion of evidential support, known as the ‘Bayes factor’. (Readers who prefer a different measure of evidential support are welcome to plug their preferred notion into my account; I expect that parallel arguments to the ones I’ll be making are available.) A Bayes factor quantifies the evidential connection between a piece of evidence $E_t$, and a pair of alternative (mutually exclusive) hypotheses. The Bayes factor of $E_t$ relative to $(H, H')$ is the ratio between the conditional probability of $E_t$ given $H$ and that of $E_t$ given $H'$, i.e., $P(E_t | H) / P(E_t | H')$. This quantifies how much more likely $E_t$ is on the assumption that hypothesis $H$ is true than on the assumption that $H'$ is true.

For our purposes, the alternative hypotheses under consideration are always a proposition and its negation, so we can define the Bayes factor of $E_t$ relative to $H$ as follows:

$$BF_{(E_t, H)} = P(E_t | H) / P(E_t | \neg H)$$

When this factor is greater than 1, conditionalizing on $E_t$ increases one’s confidence in $H$ relative to $\neg H$, with the increase in proportion to this factor. Thus, if one starts out $n$ times more confident in $H$ than $\neg H$, then, after updating one’s belief in response to evidence $E$, one should be $BF_{(E_t, H)} \times n$

---

7 Note that one of these components without the other would not give us what we want: the true theory of fundamental physics has objective probability 1 conditional on the proposition that Trump lost the 2020 election, but the fact that Trump lost the election did not mean/indicate the truth of the theory because it provided no evidence for it. Moreover, that someone is a woman does not mean/indicate that they are pregnant, despite the fact that their being a woman greatly increases the probability that they are pregnant (and so, in that sense, evidentially supports the proposition that they are pregnant).
times more confident in H than ¬H. For example, a cloudy sky bears an evidential connection to the proposition that it will rain soon, because the probability that the sky is cloudy given that it will rain soon is significantly higher than the probability that the sky is cloudy given that it won’t rain soon. Note that, although this evidential connection is reasonably strong, the informational connection between a cloudy sky and the proposition that it will rain soon is not very high because the conditional probability of rain given a cloudy sky is modest.

My notion of ‘informational connection’ is, unlike Dretske’s notion, non-factive. Even if a token of R is good evidence for p, and p highly probable given a token of R, it is possible for R to be tokened without p being true. There is still, however, a connection between degrees of informational connection and accuracy: we don’t expect to see pervasive misinformation. Representations will rarely be tokened when propositions they are informationally connected to are false.

Central to the above definition of information is the notion of probability. Corresponding to the notions of subjective and objective probabilities, there will be subjective and objective notions of information. For us, the relevant notion of informational connection must be based on objective probabilities. An appeal to a subjective notion in this context would not only make semantics problematically observer-dependent, but also compromise the reductive credentials of the resulting account since we would be relying on intentional mental states (namely, credences) in an account of intentionality. In what follows, I always use ‘informational connection’ under its objective construal.

The objective notion of an ‘informational connection’ will be the backbone of my account of content. The strength of the informational connection between a representation and a proposition is, by my lights, indicative of how ‘well placed’ that proposition is to be the content of the representation. The strength of an informational connection will not be all that matters to content determination, however. On the view I will defend, the content for a representation is not determined by that representation’s informational profile (the strength of its informational connection to each proposition), but rather by general patterns in the informational profiles of all the representations in a ‘representational system’. In other words, the recipe for getting meanings from informational connections will be holistic: we cannot determine the correct interpretation for a single mental sentence without taking into account how this interpretation fits within a larger ‘interpretation function’ for the entire system of representations. Roughly: the correct interpretation will be given by the best summary of the overall pattern of informational connections involving a structured representational system. Before developing this in detail, let me explain what I mean by a representational system and an interpretation function.

---

8 This sort of holism is one of the most significant differences between my account and other recent informational/ causal accounts in the literature (e.g., Rupert (1999), Usher (2001)). These accounts also differ in other ways: Rupert’s notion of tracking is closely tied to causation, and Usher appeals to a measure of informational strength that is importantly different from mine.
2.2 Interpretation Functions

I’m after an account of intentional content which fits well with the representational theory of the mind (RTM). The central commitment of RTM is that thoughts and mental attitudes are realized by relations to certain kinds of mental entities: mental representations. For example, if S believes that \( p \) it is in virtue of bearing a certain computational relation \( Bk \) to a mental representation that has content \( p \).

On the standard construal of RTM (Fodor, 1975), mental representations are assumed to form an internal representational system akin to spoken and written languages. Broadly construed, a representational system is any set of entities which can be specified in terms of a set of atomic symbol types, and some rules for forming complex symbols out of the atomics via one or more building operations/relations. More specifically, a representational system is any set of entities which can be picked out by a description of this form:

\[ L \text{ is a set such that:} \]

(i) All entities of atomic types \( X, Y, Z \)… (where, intuitively, \( X, Y, Z \) stand for the syntactic properties that will individuate our basic symbols.)

(ii) All symbols which can be constructed out of symbols of \( L \), in accordance with the rules: \( R_1, R_2 \)…

The idea behind RTM is that each atomic symbol can be realized in the mind by some syntactic property, in principle specifiable in neural and/or computational terms.\(^9\) For example, the primitive mental predicate DOG might be realized by a complex property of populations of cells in the nervous system, which is instantiated by any representational structure that involves that predicate (i.e., every thought or attitude concerning dogs). These primitive symbols can then be used to build complex representations, via syntactic operations realized by further neural properties/relations.\(^{10}\)

My account is in principle compatible with a wide range of representational systems, including systems that do not have subject-predicate structure, or any logical notions. For example, a system of graphs or of images can too be the target of interpretation. For illustrative purposes, it helps to consider a concrete model of what a mental language might be like.

Let \( M \) (for ‘mentalese’) be a system with the following atomic symbols and formation rules:

---

\(^9\) I don’t assume that syntactic properties are intrinsic: whether a particular entity has a syntactic property may depend on how it functions within an entire cognitive system.

\(^{10}\) Of course, nobody knows exactly how the building of complex symbols out of simple ones is achieved within the brain. But at least we know how it can be achieved by structured complexes of electrical circuits, which is a powerful argument for the idea that it is in principle achievable by physical systems consisting of a large number of simple units that have stable input-output patterns.
Atomic Symbols:
(i) Names: a, b, c, d…
(ii) Predicates: DOG, CAT, TABLE, ELECTRON, …
(iii) Logical Vocabulary: ~, &, …

Formation Rules
(◦ represents a syntactic building relation akin to concatenation)
(i) Predication Rule: If F is a predicate of M, and x is a name of M, then any structure of the form F◦x is a formula of M.
(iv) Negation Rule: If φ is a formula of M, then any structure of the form ~◦φ is a formula of M.
(v) Conjunction Rule: If φ and ψ are formulas of M, then any structure of the form φ◦&◦ψ is a formula of M.

Language M illustrates an important aspect of human thought: the syntactic rules that specify how complex representations can be formed out of simpler ones are recursive. Each one defines an operation for generating a new formula of M from any formula of M. These rules are able to define an infinite class of (syntactically distinct) formulas in a particularly concise manner. This observation will play an important role later on: we will be able to exploit the recursive nature of the syntax to concisely capture the meanings for an infinite class of formulas by means of an interpretation function that is finitely specifiable.

Given any representational system, we can define the notion of a ‘candidate interpretation function’ for that system. Candidate interpretation functions are all the possible ways of mapping expressions in a language to entities in the world. Among all the candidate interpretation functions, a select few will stand out as theoretically privileged: the mappings that assign each expression its actual content. I call these ‘adequate interpretation functions’. When vagueness/indeterminacy isn’t at issue, I will speak of ‘the adequate interpretation function’. But keep in mind that this is loose talk: what I have in mind is a range of equally good interpretation functions that largely agree on truth conditions.

Given the simple model of mentalese sketched above, I expect that adequate interpretation functions will be definable by a conjunction of clauses of this sort:

Names:
(i) I (A) = a, I (B) = b,…

Predicates:
(ii) I (DOG) = the property of being a dog…
Predication:
(iii) For all x, if x has syntactic form y ⊕ z (for predicate-type y and name-type z), then I (x) = the proposition that attributes property I (y) to object I (z).

Logical Composition:
(iv) For all x, if x has syntactic form ~ ⊕ φ (for some sentence-type φ), then I (x) = the negation of I (φ).
(v) For all x, if x has syntactic form φ ⊕ φ (for sentence-types φ, ψ), then I (x) = the conjunction of I (φ) and I (ψ).

To keep things simple, I did not include indexicals or demonstratives in this simple model of mentalese. The adequate interpretation function(s) for a language with indexicals and demonstratives are a little more complex: they map representational types and contexts to contents. As an example, consider the concept NOW. Rather than specifying a fixed semantic value for the type, we should have the interpretation map NOW together with some contextual parameters to a time, e.g.,

If x is of atomic-type NOW, I (C, x) = t iff t is the value of the time parameter in context C.  

2.3 Summarizing Informational Profiles

My account takes interpretation functions to be tools for summarizing patterns of informational connections throughout a particular epoch. These summaries can be of two kinds: they can be summaries of the informational profile for a single person, or they can summarize an entire population with a common system of mental representation. Here I will focus primarily on summaries for a single person, but I will be assuming that these two kinds of summaries will roughly coincide.

Let a ‘summary’ of the informational profile of S’s representational system during epoch E be an ordered pair (I_{def}, G) such that:

(i) I_{def} is a finite definition of an interpretation function I for S’s representational system, and
(ii) G is a generalization that exploits this mapping, which has the form:

---

11 The notion of an informational connection defined before (§2.1) doesn’t straightforwardly apply to sentence-types whose meanings vary across contexts. However, a straightforward generalization of the notion is available. Suppose sentence X varies in meaning with respect to contextual parameters: x, y, z. Taking these parameters as random variables, we can define a random variable: P (I_{x,y,z} (X) | S tokens X in context x,y,z) (and similarly for the other conditional probabilities). Given a probability measure over x, y, z, we can define expected values for these random variables, which allows us to generalize the notion of informational strength.
The informational connection between an arbitrary representation \( R \) tokened by \( S \), and that representation’s content under \( I \), is likely to be high.

Note that this generalization involves the vague term ‘high’, and therefore does not have precise truth conditions. This is so by design: a moral from the discussion of Dretske’s account is that there isn’t any specific degree of informational connection that we should demand here. Some representations bear strong informational connections to their contents, others do not. The above generalization simply invites us to have a defeasible expectation in strong informational connections between representations and their contents, without specifying any particular degree.

In the case of human beings, I will be tying the notion of tokening to belief. Thus, for \( S \) to token \( R \) at time \( t \) is for \( S \) to bear the computational relation \( Bel \) to a mental representation of type \( R \). (I suspect that this account will extend well to the contents of perception, but I won’t be exploring this issue here.)

The reason for focusing on belief is two-fold. First: I am reasonably optimistic that beliefs are reasonably well-correlated with their contents, but I’m not at all optimistic that other propositional attitudes are (consider imaginings or wondering). Second: assuming RTM, the task of giving truth-conditions for the kinds of representations that can be believed will take us a long way, since it is plausible that the same representational types will realize the other propositional attitudes. Given this, I will be understanding the above generalization as follows:

\[ G_S \] The informational connection between an arbitrary representation \( R \) believed by \( S \), and \( R \)’s content under \( I \), is likely to be high.

I propose that summaries be evaluated by how well they satisfy several competing desiderata:

(i) Fit
(ii) Simplicity
(iii) Informativeness

The idea that interpretation functions should be evaluated by the lights of several competing desiderata is not new. In response to Putnam’s model-theoretic argument for semantic indeterminacy, Lewis (1983, 1984) defended the idea that the correct interpretation for an individual is the one that satisfies two competing desiderata: how well it fits patterns of use and how eligible are the meanings it assigns to the atomic symbols. My account can be seen as one way of fleshing out this schematic proposal: the patterns of informational connections would correspond to a notion of ‘use’, and the simplicity desideratum would correspond to Lewis’s naturalness desideratum. What makes my account importantly different from Lewis’s is that I will be appealing to a purely probabilistic notion of fit.
**Fit:** I take the fit of a system to be the average strength of the informational connections between representational types and their interpretations under the system in question. Now, in the course of averaging, not all mentalese sentences need be weighted equally. Mentalese sentences that are never tokened or only very infrequently tokened (e.g. due to complexity) should count for less.

Let \( I \) be the interpretation function for the system under study and let \( M \) be the set of mental representation types that are in \( S \)'s cognitive repertoire throughout epoch \( E \). The procedure for calculating \( I \)'s fit relative to \( S \), \( E \) is as follows:

i) For each representational type \( R \) in \( M \), calculate the strength of the informational connection between \( R \) and \( I(R) \) for subject \( S \) during epoch \( E \), by averaging the strength of the informational connection between \( R \) and \( I(R) \) for \( S \) across times in \( E \).

ii) Assign a weight \( w_R \) to each representational type \( R \) in \( M \) depending on how frequently it is expected to be tokened/believed by \( S \) within \( E \) (where the expectation is based on objective probabilities).

iii) Average the informational connections for the various representational types in \( M \), weighting the strength of the informational connection for each type \( R \) (calculated in (ii)) by \( w_R \) (calculated in (iii)).

There’s an important caveat here: on the standard definition of conditional probability, the strength of an evidential connection is sometimes undefined. Consider, for example, any contradictory statement in our representational system. If \( p \) is the meaning of \( P \), then we would like our interpretation function to map \( P\&\sim P \) to the contradictory proposition \( p \text{ and } \sim p \). But what is the strength of the evidential connection between \( P\&\sim P \) and \( p \text{ and } \sim p \)? This will depend on two quantities. First, the conditional probability of \( p \) and \( \sim p \) given the tokening of \( P\&\sim P \), which is 0. And, second, the strength of the evidential connection between \( P\&\sim P \) and \( p \text{ and } \sim p \), which is given by \( P \text{ (} P\&\sim P \mid p \text{ and } \sim p \text{)} / P \text{ (} P\&\sim P \mid \sim(p \text{ and } \sim p) \). But the numerator of this expression is undefined (it involves division by 0), so the strength of the evidential connection is undefined. Similarly, in the case of tautologies, the denominator of the second component of informational strength is undefined.

Thus, the account requires that we treat propositions with extreme objective probabilities (1 or 0) differently from nomologically contingent propositions. I will take the overall fit of an interpretation function \( I \) to be the average informational strength of beliefs that are assigned non-extreme probability. The averaging procedure for evaluating fit of an interpretation function I described

---

12 Note that this issue is not specific to my version of informational semantics. Suppose I put special glasses on my cat, and as a result she comes to believe that the moon is mouse-shaped. Standard versions of informational semantics seem to imply that (under normal/channel conditions) *the moon is mouse-shaped* has objective probability 1,
above should be restricted to representational types that don’t get assigned contents with probability 0 or 1, and the generalization $G_I$ should be adjusted accordingly:

$G_I$: The informational connection between an arbitrary belief tokened by $S$, and that belief’s content under $I$, is likely to be high (provided that content does not have probability 1 or 0).

**Simplicity**: The simplicity desideratum requires that the interpretation function be concisely specifiable in a suitable vocabulary. Given the assumption that the domain of the interpretation function is infinite (as there are infinitely many mentalese sentences), the simplicity constraint rules out mappings that can only be given by an infinite list of each mentalese sentence and its interpretation. An interpretation function that can be defined recursively, on the other hand, will allow us to keep the summary simple enough.

Given the assumption that highly unnatural/disjunctive predicates count against simplicity, simplicity considerations will favor interpretation functions that map our atomic concepts to natural properties/relations over ones that map them to less natural properties/relations. As we will see, this will be important in explaining how to avoid the kind of indeterminacy that other versions of informational semantics face.

**Informativeness**: Summaries of informational connections are better to the extent that they tell us more about how a subject’s mental representations are correlated with features of the environment. A preference for informative summaries helps explain why interpretation functions that only assign meanings to a few mental representations are not good, despite being simple and accurate. Similarly, the informativeness desideratum is needed to rule out certain trivial interpretation functions: for example, the function that maps each representation $R$ to the proposition that $R$ is tokened at the relevant time will be very simple and have good fit, but is not at all informative.\(^\text{13}\)

Some readers may find it helpful to think of this account as a new kind of ‘interpretationism’. Broadly speaking, interpretationism takes semantic facts to fall out of the most ‘charitable’ overall interpretation of a subject’s representational system. In the original formulation of the view (Davidson, conditional on my cat’s having this belief. How could this be, if the moon’s being mouse-shaped is nomologically impossible? This belief has to be somehow excluded from the scope of the account.

\(^\text{13}\) For the purposes of this paper, an intuitive understanding of informativeness will do. But, eventually, I hope we will be able to make this notion precise along the following lines. Suppose I start out knowing nothing (or very little) about how a subject $S$ is embedded in its environment, but I have a method for working out what mental state they're in at each time (where that state is described syntactically). If I let a summary $\langle I_{def}, G_I \rangle$ guide my expectations about the state of the environment (i.e., if I form a credence function which sets the conditional probabilities as $G_I$ invites me to), how well informed will I be about the state of the environment? The more informed I am, the more informative $\langle I_{def}, G_I \rangle$ will have been.
1973, 1974), charity was tied to truth: a charitable interpretation is one that makes as many of the interpreted person’s utterances (and beliefs) come out true as possible. Other interpretationist approaches (Lewis, 1974; Williams, 2019) have de-emphasized the role of truth, holding that charitable interpretations should aim instead for the rationality of the interpreted person’s mental states. This shift is motivated, among other things, by the idea that an unlucky subject who happens to be surrounded by lots of misleading evidence is most charitably interpreted as having lots of false beliefs. My approach is, in a sense, intermediate between these two. I maintain that interpretation functions aim to make beliefs come out as strongly informationally connected to their contents. This enables me to get the prediction that subjects in very unlucky circumstances are sometimes best interpreted as having mostly false beliefs, without appealing to normative notions like rationality which are already in need of naturalistic reduction.

Before going on to defend the account, let me clarify its intended scope. While the primary targets of my account are human propositional attitudes, it can be applied to any kind of representational system, no matter how simple. In the absence of any criteria for deciding when some physical variables realize a system of representations, it might seem like my account will lead to pan-representationalism: every sufficiently complex physical system has a set of best summaries for its informational profile.

Perhaps the internal states of some things won’t have informational profiles that lend themselves to being efficiently summarized (e.g., rocks), but some non-mental things certainly will have good summaries. To take a common example, suppose we take the rings in a tree trunk as a representational system. My account predicts that there is a correct interpretation function for the tree—namely, the function that maps the number of rings at a time to the tree’s age at that time.

I don’t deny that there is an intuitive distinction between the tree rings and the kinds of mental representations that make up beliefs and desires. However, I don’t think that an account of how representations get their meanings needs to explain this distinction. Following William Ramsey (2010), I distinguish two tasks:

(i) Understanding how the meaning of a mental representation gets fixed.
(ii) Understanding what it is for some feature of reality to function as a representational vehicle.

My focus in this paper is exclusively on project (i), but I think the distinction between tree rings and human beliefs will come when we have a handle on (ii). There will be some story to tell about why the tree rings don’t function as representations. So, despite the fact that their informational profile has a nice summary, we will have some independent reason to deny that they have mental states. This is not to say that I endorse a purely non-semantic criterion for deciding what counts as a representation, or a mental representation. The second project may end up being intimately tied to the first. My own view is that a physical variable functions as a representation when its adequate interpretation function(s) is/are invoked by the best explanations of an associated pattern of behavior. I’m quite
confident that the behavior of trees is best explained without citing any interpretation functions, so I expect that we will be able to draw a principled distinction here.\footnote{The account this paper develops is compatible with other ways of approaching (ii). For example, someone could combine my account of content with an account of the notion of a representational vehicle which appeals to their biological function and/or their role in computational processes.}

2.4 Which probabilities?

We saw before that Dretske’s account faces the question: where do objective probabilities come from? Following Dretske, I take probabilities to be associated with lawlike regularities. But not just any conception of the relation between laws and probabilities will do.

On one conception of objective probability — the ‘dynamical conception’ — all objective probabilities come from the dynamical laws of fundamental physics. These dictate how likely a total physical state of the world is, conditional on a previous state. On this view, all objective probabilities are conditional in form, and only apply to very specific pairs of propositions — propositions about total states of the world at a time. There isn’t, on this view, a single correct objective probability distribution over all propositions. If we ask for the probability that Obama will eat Cheerios for breakfast tomorrow, there simply isn’t an answer. There also isn’t a determinate probability that Obama will eat Cheerios for breakfast tomorrow, conditional on the information that he bought a box of Cheerios this morning.\footnote{There are many possible physical states of the world that realize Obama’s buying a box of Cheerios this morning. In order to get a probability for his eating Cheerios tomorrow, we need to know how likely each of these possible states are, and the dynamical laws don’t tell us this.}

If we adopted the dynamical conception of probabilities, my informational account of content would be inadequate at best, and incoherent at worst. For this account appeals to conditional probabilities of propositions involving representational states given propositions about worldly states (and vice versa). But no reason to think such pairs of propositions have well-defined dynamical conditional probabilities. In this section I present an alternative way of thinking about the relevant probabilities that fits well with my account. This will serve two purposes. Firstly, the availability of such a notion of probability shows that my account is on firm metaphysical footing: it doesn’t appeal to more structure than the world offers. Secondly, having a concrete conception of probability in mind will be useful to establish that the account makes the right predictions about cases (§3).

In other work, I’ve argued that we should think of the probabilistic generalizations of the special sciences as concise summaries of the frequency patterns throughout our ‘modal neighborhoods’ (Gómez Sánchez, 2020). A modal neighborhood is a set of nearby possible worlds, where each world in the set (i) respects the actual laws of physics, and (ii) is physically similar to the actual world at some
time in an epoch of interest E. Such a neighborhood includes all the worlds where things went a little differently from the way they went in our world at some time in the epoch. (For example, the closest worlds where you decide to wear a different outfit today are in the modal neighborhood, and so are the closest worlds where you spill a beverage later today). How far in modal space the neighborhood we end up summarizing extends may vary from one scientific field to another. But, in general, we aim to cover the largest modal neighborhoods that can be covered with simple, informative and accurate theories.\(^{16}\)

I call the statements that feature in the best summary of our modal neighborhood(s) ‘crystallized regularities’. I contend that, in our world, the crystallized regularities will include—besides the physical laws—a number of evolutionarily contingent regularities about our environment, including ones about our cognitive architecture and our system of mental representation (Gómez Sánchez, 2020).

In line with this conception of the special sciences, I suggest that we think of interpretation functions and their associated generalizations as summarizing frequencies throughout our modal neighborhood. Let \(M\) be an independently motivated probability measure over physical possibilities. We can define relative modal-neighborhood frequencies for arbitrary pairs of propositions as follows:

\[
F(p \mid q) = M(\text{nearby } p \& q\text{-worlds}) / M(\text{nearby } q\text{-worlds})
\]

For the purposes of this paper, I will take these modal frequencies to play the role of probabilities: they will serve as the objective basis for informational relations between mental structures and worldly states. (It is an open question whether these modal neighborhood frequencies are apt to play other roles that objective probabilities play.)

The aforementioned conception of special science laws invites a further development of the present account: we can take adequate interpretation functions and their associated generalizations to be special science laws. In the case of humans, the envisaged law would be of the form:

\textbf{G}_{1\text{-humans}}: For any human S, the informational connection between an arbitrary belief tokened by S, and that belief’s content under I, is likely to be high (provided that I(S) does not have probability 1 or 0).\(^{17}\)

\(^{16}\)I take distances between worlds to be objective, and based on the world’s fundamental physical structure. I assume that fundamental physics grounds relatively determinate distances between physical states of the world at a time. For any two worlds \(w, w'\) that share the laws of physics, I define the distance between them, relative to epoch E, as the degree of similarity of their physical states at time \(t\), where \(t\) is the time within E at which their states are most similar.

\(^{17}\)Note that, unlike the summaries I’ve been focusing on in this paper, this would correspond to a population-wide summary: it would summarize the informational profiles of everyone’s mental states at once. This population-level summary assumes that there are non-semantic criteria for identifying representations in different people’s minds as being of the same type.
Even assuming the view that special science laws are summaries of some kind, establishing the lawhood of $G_i$-humans requires much more than I can accomplish here. Earlier, I stipulated that I would be focusing on summaries that take a certain form (interpretation functions coupled with particular kinds of generalizations), and I also stipulated that they would be evaluated in accordance to simplicity and fit under particular conceptions of those virtues. It remains to be shown that the best summary of the informational profile for our mental language (in this stipulated sense) is also a part of the best general summary for our modal neighborhood (in the sense of ‘best’ that is relevant to lawhood). While this strikes me as a plausible conjecture, and worthy of further investigation, the account I've given in this paper does not depend on it. My primary goal here is to propose necessary and sufficient conditions for content in terms of information which give plausible verdicts on a wide range of cases. A proper understanding of how the defined notion fits into the nomic structure of our world will have to await future work.

3. Verdicts

Now that the full account is on the table, I discuss how it handles paradigmatic cases of human representation, including those that are usually considered beyond the scope of informational semantics. My goal is to make it plausible that—under the intuitively correct interpretation of mentalese—many of our beliefs bear reasonably strong informational connections to their contents, and that non-standard interpretations of our systems of mental representation are either too complex, or worse in overall fit, etc. As I discuss these cases, I will also seek to make it clear how (and to what extent) misrepresentation is possible and indeterminacy is avoided.

3.1 Observational Terms

To start, consider cases of human belief that informational approaches are reasonably placed to handle: beliefs involving terms that concern objects and/or properties that are easily tracked on the basis of their interactions with our sensory organs (without recourse to too much background knowledge). Some examples of highly observational representations are concepts like GREEN, TREE, ROUND, HOT, CAT, COFFEE, RIGID, MITTENS. In what follows I will work through a paradigmatic example of an observational mentalese sentence in some detail. I hope it will be clear that the arguments extend to almost all observational sentences.

Take the proposition that Bernie Sanders wore mittens on Jan 20th, 2021 (Bernie, for short). Consider an arbitrary human subject S that believes this in virtue of tokening a mental representation BERNIE.

---

18 I'm using the term 'concept' here in a syntactic sense: a concept is any representation-type which can be a constituent in a belief. (Fodor, 1998)
By relying on the connection between probabilities and modal frequencies explained in §2.4, I will argue that interpretation functions that assign BERNIE the content Bernie will be better than salient alternatives, given the criteria of fit and simplicity.

To evaluate the informational strength of the connection between BERNIE and Bernie (for subject S during epoch E), we must estimate the following two quantities:

(i) \( P(Bernie \mid Bel(S, BERNIE)) \)
(ii) \( P(Bel(S, BERNIE) \mid Bernie) / P(Bel(S, BERNIE) \mid \sim Bernie) \)

Consider (i): take the nearby worlds where S believes BERNIE.\(^{19}\) In what proportion of these worlds does Bernie obtain? Given what we know about how people form beliefs in worlds like ours, their belief in these worlds will probably have been caused in one of two ways: (a) by some perceptual stimulus that is likely to have been caused by Bernie wearing mittens, or (b) a chain of testimony that is most likely to have originated from a perceptual encounter with Bernie wearing mittens. In either case, the belief is most likely to have been caused by Bernie wearing mittens. So \( P(Bernie \mid Bel(S, BERNIE)) \) will surely be greater than 0.5, and possibly significantly greater.\(^{20}\)

Now consider (ii). For typical subjects, the frequency with which they believe BERNIE is greater among nearby Bernie-worlds than among nearby \( \sim \)Bernie-worlds. So \( P(Bel(S, BERNIE) \mid Bernie) \) will be significantly greater than \( P(Bel(S, BERNIE) \mid \sim Bernie) \). (Note that this is compatible with the claim that in many nearby worlds where Bernie wears mittens, hardly anyone notices.) Putting these observations together, we can conclude that there tends to be a strong informational connection between BERNIE-beliefs and the proposition Bernie. This makes Bernie a good candidate content for BERNIE: interpretations that assign BERNIE this content will gain points in fit for that sentence.

Note that the above reasoning assumes that our perceptual systems and our ways of responding to linguistic stimuli are somewhat reliable throughout worlds that are similar to ours in their physics and macro-structure, but it does not assume that they are infallible under some specified ‘normal’ or ‘channel conditions’. The view can accommodate infinitely many nearby worlds where S’s perceptual system is not malfunctioning, but S forms the belief BERNIE on the basis of incorrect testimony, a

\(^{19}\) What should count as a ‘nearby’ world in this context can be decided as follows: take the largest neighborhood possible, without going far enough to start seeing violations of the regularities of special sciences (e.g., biology and neuroscience) that we deem to be robust (non-accidental, counterfactual supporting, and explanatory). This may seem circular: I’ve said before that the robustness of special science regularities will depend on whether they hold throughout our modal neighborhood. But this desired constraint can be reformulated to avoid circularity: construct systems of generalizations for modal neighborhoods of different sizes, and then choose the biggest modal neighborhood that has a sufficiently informative system.

\(^{20}\) We also need to consider situations in which a belief is formed on the basis of an episodic memory of a perceptual encounter. Because episodic memory is a much noisier source of information than perception is, such beliefs are likely going to more often be held falsely. I doubt that this will make a difference to the overall point that observational beliefs are more often true than false.
misleading percept or an inaccurate memory. The view can also accommodate possible situations where S forms the belief on the basis of wishful thinking, or in some entirely irrational manner. 21 All that is required for a strong informational connection is a certain kind of modal pattern: more often than not, the belief is true when held, and the belief is much more likely to be held when its content is true than when its content is false.

Now, suppose that S has a robust tendency to mistake big gloves for mittens when it’s dark. The conditional probability of Bernie wore mittens-or-big-gloves-in-the-dark on Jan 20 th given that S tokens BERNIE will be higher than the conditional probability of Bernie given that S tokens BERNIE. (In fact, this follows from the axioms of probability alone). Moreover, if the mistake is systematic, then BERNIE may be as good evidence for Bernie wore mittens-or-big-gloves-in-the-dark on Jan 20 th as it is for the proposition that Bernie wore mittens on Jan 20 th . However, an interpretation that assigns the meaning mittens-or-big-gloves-in-the-dark to MITTENS will plausibly be worse in virtue of assigning MITTENS a highly unnatural property.

Next, consider the proposition that Bernie went outside on Jan 20th. Let us grant, for the sake of argument, that in every nearby world in which Bernie goes outside on Jan 20th he wears mittens, and in every world where he doesn’t go outside on Jan 20th he doesn’t wear mittens. Given this, we get the same modal pattern as before: (i) Bernie is likely to have gone outside on Jan 20th conditional on S believing BERNIE, and (ii) BERNIE is much more frequently believed by S in worlds where Bernie went outside on Jan 20th than in worlds where he didn’t. Given the stipulation that the nearby Bernie-worlds coincide with those worlds where Bernie goes outside on Jan 20th, these propositions will be matched with respect to the two components of informational strength. It follows that the informational connection between BERNIE and Bernie went outside on Jan 20th will be just as strong as that between BERNIE and Bernie.

Despite this, my account is able to explain why interpretation functions that map WEARS MITTENS to goes outside are worse. The holistic approach to measuring fit is what does the work. Consider beliefs involving representations of the form ~WEARS MITTENS (X, t). The modified interpretation function would presumably map such representations to propositions of the form X doesn’t go outside at t. But these propositions would not be strongly informationally connected to the representations mapped onto them. Of course, we could modify the interpretation function further (e.g., by altering the meanings assigned to ‘~’ or to proper names), but there’s no reason to suppose that this can be done systematically, without disrupting further informational connections.

21 However, this account cannot accommodate reliable misrepresentation in the sense of Mendelowici (2012). If, by the lights of interpretation I, people are always (or almost always) mistaken about their beliefs involving some concept X in actual and nearby worlds, then I will probably lose to another interpretation function that interprets X in some other way.
Similar reasoning shows how interpretations that map BERNIE to its proximal causes can be excluded. Suppose that S1, S2, ... , Sn are all the possible patterns of sensory stimulation that lead a person to form the belief BERNIE. (For example, S1 might be a pixel-by-pixel description of a photo of Bernie wearing mittens at the presidential inauguration). Then the disjunction $S_1 \vee S_2 \vee \ldots \vee S_n$ will be strongly informationally connected to BERNIE. But it is not at all plausible that we can construct a simple compositional mapping from parts of the sentence BERNIE to aspects of the world which will support this content assignment (much less one with good overall fit).

Finally, consider disjunctive/gruefied properties which are well tracked by some concept. Consider the property of being a mitten*: to be a mitten* is to be a mitten or a glove made out of exactly $n$ atoms. This property is clearly not one we think about when we think about mittens. But gloves made out of exactly $n$ atoms are extremely rare, so the instances of mittens and mittens* in nearby worlds will coincide almost perfectly. Because of this, there will be a strong informational connection between BERNIE and the corresponding proposition about mittens* (just as strong as the informational connection between Bernie and BERNIE). Moreover, the same can be said for any other mental sentence involving MITTENS. What, then, favors mittens over mittens*? The answer to this question must, I think, invoke the difference in naturalness between the property of being a mitten and the property of being a mitten*. Definitions of interpretation functions that cite unnatural properties are highly complex, and therefore unlikely to belong in the best summary of the informational profile of our representational system.

Note that mittens and mittens* effectively compete for a single slot in the correct interpretation function. For any given system, only one of them can be the value of the interpretation function for MITTENS. Since it is plausible that interpretation functions that differ only by which of these two properties is the value of $I$ (MITTENS) will be the same with respect to fit, simplicity must break the tie. A similar story can be told for Quinean alternatives: rabbit and undetached rabbit part may be equally well tracked by the concept RABBIT, but interpretations that map RABBIT to rabbit will be favored because of their greater simplicity.22

Here I’ve been focusing on a single proposition, Bernie. But I hope it is clear that the style of argument I’ve given generalizes to a wide variety of observational beliefs, including: X IS GREEN, X IS ROUND, X IS A CAT, X IS A CAT IS ON THE MAT, etc. This story may not apply to all observational beliefs. It won’t apply, for instance, to the belief that a magician has just caused a bird to disappear or teleport,

---

22 We’ve seen so far that my account is able to yield reasonably determinate contents for observational beliefs. I should note, however, that my account also accommodates indeterminacy in places where we expect to see it. Take our mental representation GREEN. No discontinuity in color space serves as a precise boundary between green and blue. Moreover, subjects don’t have stable dispositions to apply GREEN to borderline shades. Hence, there will likely be various interpretation functions that assign slightly different meanings to green but are equally good with respect to both fit and simplicity. If so, then it is indeterminate which one is best, and so indeterminate which gives the correct meaning.
for these beliefs are never correctly held in nearby worlds (see also the example in fn. 12). Such beliefs can nonetheless have determinate contents, by having constituents that feature in many other beliefs that do bear strong informational connections to their contents.

3.2 Theoretical and logical terms

It is often recognized that informational and causal accounts are most plausible for perceptual states and observational beliefs. A background assumption has been that the correct account of content will proceed in a piecemeal fashion: first, we give a causal/informational story for observational concepts and some other story for logical vocabulary (perhaps a restricted version of conceptual role semantics (Block, 1986)); once we have meanings for this initial fragment of mentalese, we give a different kind of story for highly theoretical terms (perhaps a sophisticated kind of descriptivism).

A potential advantage of my version of informational semantics is that it may extend to at least some non-observational terms. While I don’t claim that my view is sure to give the correct verdicts for all non-observational concepts, I think my account is at least approaching a general account of content. I think the account plausibly covers logical vocabulary, and at least some highly theoretical concepts.

Consider the concept ELECTRON. Set aside deferential uses of the term (which I’ll discuss in a moment), and let us restrict our attention to subjects whose use of the concept ELECTRON is based on a full understanding of a theory in which this concept is embedded. Call this theory T.

With regards to electrons, the kinds of mental sentences that are frequently tokened are not of the form ‘X IS AN ELECTRON’. (I expect that circumstances under which someone can latch onto an individual fundamental particle in thought are rare.) While I’ve been heavily relying on these kinds of singular beliefs to illustrate how the account works, such beliefs don’t have any kind of special status on my view. In fact, scientists who are competent with theory T will have many other kinds of beliefs that are informationally connected to facts about electrons: e.g., beliefs about the positions and velocities of electrons that they are able to measure, beliefs that large numbers of electrons passing through a lightbulb’s filaments are causing some visible glow, beliefs about how their pressing a key on their computer keyboard affects the electrical flows within the computer’s logic gates, affecting what they see on their screen, etc. These informational connections at least make it plausible that the intuitively correct interpretation function (which maps ELECTRON to electron) will fit better than interpretations that map ELECTRON to a macroscopic property (like green), or other physical properties (like proton or mass). And, given how natural the property electron is, this may suffice to lock ELECTRON onto it.

---

One potentially strange aspect of this story is that the theory T with respect to which ELECTRON is defined plays no direct role in giving ELECTRON its meaning. In fact, since the intuitively correct interpretation function plausibly maps T to a proposition that has probability 1 or 0, the strength of the informational connection between beliefs in T and T’s content will be undefined. So the story that I’m telling crucially depends on subjects having the ability to use the theory T to form further beliefs in contingent propositions involving the property electron. Is it not possible to have a theoretical concept without such further beliefs? I suspect that it is, which would reveal a limitation in the purely informational notion of fit. Perhaps a more sophisticated notion of fit is needed to cover a wider range of theoretical concepts.

Finally, consider logical vocabulary. It is standardly thought that logical notions aren’t even in the purview of informational semantics. One of the strongest aspects of my view is that it can arguably handle logical notions just as well as observational ones.

Suppose that Mentalese has mental symbols for negation and conjunction, ‘¬’ and ‘&’, and that they behave syntactically like negation and conjunction do in a standard formal language. The interpretation function must map arbitrarily complex sentences involving ‘¬’ and ‘&’ to propositions which they are strongly informationally connected to. Arguably, this is best achieved via recursive clauses in the definition of our interpretation function I, appealing to logical relations between propositions (like we saw in § 2.2):

For all x, if x has syntactic form ¬φ (for some sentence-type φ), then I (x) = the negation of I (φ).

For all x, if x has syntactic form φ&ψ (for sentence-types φ, ψ), then I (x) = the conjunction of I (φ) and I (ψ).

That this makes for well-fitting interpretation functions is extremely plausible. Take conjunction. If mental representations of type φ are well informationally connected to p, and mental representations of type ψ are well informationally connected to q, it is hardly surprising that φ&ψ will be informationally connected to p and q; after all, people tend to believe φ&ψ when they also believe each conjunct, and p and q is true when each conjunct is true. Of course, similarly good fit can be achieved by gerrymandered functions in the vicinity of conjunction (e.g., a function that maps all pairs of propositions to their respective conjunctions except for a random pair, which it maps to one of the conjuncts). Once more, this interpretation function will be ruled out because it is significantly more complex than the standard interpretation.
3.3 Deference to experts

I want to briefly discuss mental representations that most people apply by deferring to experts. Classic examples of this include notions like ‘elm’ or ‘arthritis’. This section explores how my account fares with respect to these concepts. I will show that, if we make some assumptions about the modal robustness of the connection between words and concepts, my account yields the right predictions: it can accommodate the special role that beliefs of experts play in fixing the content of such representations, without needing to give a separate account for these sorts of representations (and without relying on the murky notion of an ‘expert’).

Let S name the subject of Tyler Burge’s famous thought experiment (Burge, 1979). S believes that she has arthritis, that she’s had it for a while, that her frequent joint pain is a symptom of arthritis. She's also quite confused about the nature of arthritis, and her confusion leads her to think she has arthritis in her thigh. Despite this confusion, S would be willing to be corrected by experts: if a doctor were to tell her that arthritis is a joint-disease and cannot afflict one’s thighs, S would trust the doctor on this and update her beliefs accordingly.

By arguments parallel to those given in the previous two sub-sections, we could argue that doctors’ beliefs about arthritis are by and large strongly informationally connected to their contents. What I will show now is that, given the informational connection between doctors’ beliefs and propositions about arthritis, beliefs about arthritis of non-experts will also satisfy the required informational condition.

Let ARTHRITIS$_D$ name the concept of arthritis that doctors have in virtue of grasping some theory T, and let ARTHRITIS$_S$ be S’s concept, which intuitively gets its meaning via the concept ARTHRITIS$_D$. Arguably, the association between the concept ARTHRITIS$_D$ and the word ‘arthritis’ is modally robust (i.e., constant throughout nearby worlds), and so is the association between ARTHRITIS$_S$ and the word ‘arthritis’. Nearby worlds, here, are worlds that we can get to by making small physical changes at one time, and there are not many small physical changes to the state of the world now that would break the association between the English word ‘arthritis’ and the respective concepts. Given this stable connection, in nearby worlds sincere utterances of the form ‘x has arthritis’ made by doctors are correlated with beliefs of the form X HAS ARTHRITIS$_D$.

I assume that, throughout nearby worlds, doctors don’t often lie when they’re talking about arthritis (who has it, what are the symptoms, treatments, etc). Since, in each nearby world, S strives to correlate her ARTHRITIS$_S$-beliefs with one or more doctors’ ARTHRITIS$_D$-beliefs, S’s beliefs involving ARTHRITIS$_S$ will be accompanied throughout nearby worlds by corresponding beliefs in the minds of one or more doctors around her. Unless S were systematically unlucky and happened to defer at each world to a doctor who is epistemically unlucky at that world, we should expect that her beliefs
involving ARTHRITIS$_S$ would come to have informational profiles that (roughly) mimic the informational profiles of doctors’ beliefs about ARTHRITIS$_D$. If so, the property arthritis should be a reasonably good candidate content for ARTHRITIS$_S$ (despite S’s confusions about its nature).

Now, as Burge notes, if S had been in a very different linguistic environment, things could have gone differently. My account delivers this prediction too. Consider a region of modal space where S is intrinsically similar to the way she is in the original scenario, but doctors associate the word ‘arthritis’ with a concept that best tracks some other disease. In a modal neighborhood like that, S’s deferentially formed beliefs will be better informationally connected to the alternative disease.  

### 3.4 Simple representational systems

If my arguments in §3.1-3.3 are correct, then we have a purely informational account of content with the resources to handle a broad range of human representations. I want to end by considering the application of this account to proto-representational systems: internal states of living beings that have evolved because of their informational connections to certain features of the environment, but which lack the rich compositional structure of our mental language. These cases illustrate that, given my account, determinacy of content is harder to achieve for simpler representational systems.

First, take an example from Dretske: bacteria in the northern hemisphere have small inner magnets (magnetosomes) which cause them to swim parallel to magnetic field lines. In their normal environment, this enables the bacteria to avoid oxygen-rich water.

If we think of the magnetosomes (and the directions in which they pull bacteria) as forming a representational system, how should we interpret it? Millikan (1989) argues that informational accounts cannot get the right results in a case like this. Her intuition is that the pull of the magnetosome at a given time represents the direction of oxygen-free water, whereas “the most reliable natural information that the magnetosome carries is surely not about oxygen-free water but about distal and proximal causes of the pull, about the direction of geomagnetic or better, just plain magnetic, north” (Millikan, 1989).

While it is true that my account does not deliver the prediction that Millikan deems correct, it doesn’t exactly give the verdict what Millikan expects informational accounts to yield. In the bacteria’s

---

24 What if the association between the word ‘arthritis’ and the concepts ARTHRITIS$_S$ and ARTHRITIS$_D$ are not modally stable? What if, that is, there is wide diversity in how words get used throughout the modal neighborhood? Then the story about deferential uses will not be as neat as I’m depicting it here. The interpretation function would have to capture a more complex informational pattern: where a single concept tracks different properties in different sub-regions of the modal neighborhood depending on how words are being used in those regions. I suspect that, to handle this, we would need an interpretation function that treats ARTHRITIS$_S$ as a context-sensitive expression (see §2.2), where the context is partly defined by how words are being used by experts. I’m not sure how to make this notion of context precise, so I leave this as an open question for future work.
environment, oxygen-free water is perfectly correlated with the magnetic north, not only in the actual world but also in almost all nearby worlds. Thus, the informational connection between the magnetosome’s direction of pull and the (egocentric) location of oxygen-free water is plausibly as strong as the connection between the magnetosome’s direction of pull and the orientation of Earth’s magnetic field. Given this, I suspect that there will be two interpretation functions that fit the informational facts equally well. Unless one turns out to make for a simpler or more informative summary than the other, there won’t be enough basis for assigning determinate meanings.

This doesn’t strike me as a cost, however. After all, the kinds of explanations that we’re inclined to give for how these bacteria interact with their environment don’t seem to presuppose reasonably determinate content-assignments to the pulls of magnetosomes. (By contrast: our explanations of human behavior do seem to presuppose reasonably determinate meanings). I doubt that ‘These bacteria are swimming in direction d because they represent d as the direction of oxygen-free water’ is a very good explanation. Here, simpler explanations seem to suffice: ‘These bacteria are swimming in direction d because d is north and their magnetosomes interact in such-and-such way with Earth’s magnetic field’. Or: ‘These bacteria are swimming in direction d because their magnetosomes pull in the direction of oxygen-free water, and there’s oxygen-free water in that direction’. Depending on the context, one or both of these explanations would be satisfactory, so an appeal to representation seems gratuitous.

Next, consider the classic example of a frog that sticks out its tongue in the presence of flies. Suppose that the frog accomplishes this in virtue of tokening a perceptual representation of the form ‘S, l’ (where, intuitively, S tracks flies and l specifies an egocentric location of the perceived fly). In the frog’s usual environment, the probability that a fly is present at a given egocentric location l is high given a tokening of ‘S, l’, and the probability of the tokening is much higher when a fly is in fact present at that location than when it is not. But these same conditions will also be satisfied for the proposition that a little-black-moving-ambient-thing is present at l or the proposition that a fly-or-bee-bee is present at l. But, given my view, there might still be a determinate answer concerning which property the frog represents.

Assuming all three properties (fly, little-black-moving-ambient-thing and fly-or-bee-bee) are equally well tracked—assuming, that is, that the interpretation functions that differ in which of these three they assign to S fit the informational profile of S equally well—fly would plausibly a better candidate meaning for S in virtue of being more natural. But, in fact, the informational connections between representational states of the form ‘S, l’ and facts about the locations of flies are not as strong as the informational connections between those representational states and facts about the locations of little-black-moving-ambient-things or fly-or-bee-bees. Fly-or-bee-bee will likely be ruled out as a candidate meaning on the basis that it is too disjunctive/unnatural. Little-black-moving-ambient-thing is not disjunctive but still quite unnatural, so it would only be selected as the meaning if this were to significantly improve fit. I suspect that moving fly will make for a better interpretation function—but

Add citation.
the exact verdict will depend on the degrees of naturalness of the various notions, and the correct way of trading off the various criteria. I see no decisive reason to expect indeterminacy here, unless there’s indeterminacy regarding how natural notions are, or how different desiderata (e.g., simplicity vs. fit) are to be balanced.

Some readers may worry that the appeal to naturalness and to trade-offs for competing desiderata leads to an insufficiently objective conception of an ‘adequate interpretation’. But these sorts of considerations are needed whenever we try to choose among theories/models in non-fundamental domains. I think naturalness facts are objective (Gómez Sánchez, 2021), and I suspect that there are objective facts about how various kinds of desiderata are to be traded off in the course of scientific theorizing. But if I’m wrong about this, the lack of objectivity will infect all of psychology—not just the theory of content.

REFERENCES


—— (2021) Naturalness by Law. Manuscript


