Representations are Rate-Distortion Sweet Spots

Abstract

Information is widely perceived as essential to the study of communication and representation; still, theorists working on these topics often take themselves not to be centrally concerned with "Shannon information", as it is often put, but with some other, sometimes called "semantic" or "nonnatural", kind of information. This perception is wrong. Shannon's theory of information is the only one we need.

I intend to make good on this last assertion by canvassing a fully (Shannon) informational answer to the metasemantic question of what makes something a representation, for a certain important family of cases. This answer and the accompanying theory, which represents a significant departure from the broadly Dretskean philosophical mainstream, will show how a number of threads in the literature on naturalistic metasemantics, aimed at describing the purportedly non-informational ingredients in representation, actually belong in the same coherent, purely information-theoretic picture.

1 Information, Shannonian and Dretskean

In what follows I will use a random variable, S, to encode the state the world is in, and another random variable, M, for signals. How should we characterize the information that values of M (i.e., individual signals) carry about values of S (i.e., individual world states)? The most basic quantity with which information theory records dependence among two random variables is the *mutual information* between them. This quantity being an expected value, Dretske (1981, p. 52f) claims, renders it unsuitable for an analysis of representational status, and it should be substituted by notions that record relations between individual states, S_i, and individual signals, M_j. The basic relation which substitutes mutual information in contemporary Dretskean accounts is that of *making a probabilistic difference* (Scarantino 2015): a signal M_j makes a probabilistic difference to the instantiation of a state S_i iff the following *basic inequality* holds:

$$P(S_i|M_j) \neq P(S_i)$$

Nearly all the accounts of information developed in the recent, and not so recent, philosophical literature on this topic are variations on, and attempts to quantify, this inequality. For illustration, in Skyrms (2010, p. 36) the "information in [M_j] in favor of [S_i]" is defined as the *pointwise mutual information* (Also *pmi* henceforth) between state and signal. There is a direct relation between pmis and the basic inequality: the former are nonzero iff the latter is true.

The running thread connecting most prominent contemporary accounts of information is that all there is to Shannon's information theory, at least for the purposes of investigating the nature of representation, is two quantities: the unconditional probability of states and the probability of states conditional on signals, perhaps rearranged as the logarithm of their ratio, or in some other way. Unsurprisingly, from this it is routinely concluded that there is much more to representation than information. This conclusion is premature: informational content in the Dretskean tradition is not by a long shot all there is to information theory. This should not be taken to imply that information is all there is to representation—for one thing, I believe with teleosemanticists (Millikan 1984; Papineau 1987) that teleofunctions have a role to play in a complete theory of representation—but it does mean that no Dretske-style "semanticized information" needs to be recognized, over and above the quantities studied in information theory proper. I will argue that it also means that some prominent proposals as to ways to bridge the information-representation gap are, in fact, unwittingly appealing to informational structure.

In the following section I review two such proposals. My aim is not to argue against them—they are built upon largely correct insights. I will instead aim at showing that a better informed understanding of information provides a way to incorporating these insights in a unified, purely information-theoretic picture.

2 Bridging Information and Representation

2.1 Many-to-One-to-Many Architectures

The first proposal is that it is not enough that representations carry information; on top of that, they must sit in the right place in a certain cognitive architecture. Sterelny (2003), for example, has argued that the emergence of representations is enabled by two prior evolutionary transitions: from "detection" to "robust tracking", on the one hand; from "narrow-banded" to "broad-banded" behavioral responses, on the other. Robust tracking is in essence a *many-to-one* relation between world state and signal: many sensory inputs give rise to one and the same representation. Other theorists have advocated similar architectural constraints on representational vehicles. Famously, Burge (2010) places a great deal of weight on *perceptual constancies* in his characterization of perceptual representation (Burge 2010, p. 413.) This is a variation on Sterelny's idea and, as such, a many-to-one architectural constraint on representational status.

As for broad-banded responses, in these systems a single representation will be flexibly dealt with, resulting in different courses of action, depending on the context where the representation is tokened. Response breadth is in essence a *one-to-many* relation between representational vehicle and output: one representation, many agential outputs.

2.2 Reference Magnetism

A second proposal has been to focus on the entities that should figure in the content of simple representations. The suggestion, typically, is that represented entities should be appropriately *natural*, or *real*. For example, Dan Ryder (2004, 2006) has argued that neurons become attuned to *sources of correlation*. These entities are closely related to Richard Boyd's *homeostatic property clusters* (also HPC henceforth, Boyd 1989): HPC theory identifies natural kinds with clusters of properties which tend to be instantiated together, and such that this frequent co-instantiation is not just a statistical fluke. What Ryder calls sources of correlation are the grounds for these HPC-related frequent co-instantiations—whatever it is that makes them *not* statistical flukes. Ryder claims that many of the representations the brain trades in target sources of correlation. Martínez (2013) and Artiga (forthcoming) have made more general cases that simple representations preferably target HPCs (Martínez), or properties that best explain the co-occurrence of other properties (Artiga).

A similar idea has been explored in an entirely independent line of enquiry starting with Lewis (1983): "among the countless things and classes there are ... [o]nly an elite minority are carved at the joints, so that their boundaries are established by objetive sameness and difference in nature. Only these elite things and classes are eligible to serve as referents" (Lewis 1984, p. 227). This is what Sider (2014, p. 33) calls *reference magnetism*.

As I show in section 4, these two ideas, although apparently disparate, are in fact closely related, and the explanatory payback they bring to representation-involving talk depends on their informational underpinnings.

3 Information Theory is a Source-Channel Theory

Philosophy has understood information theory as a mostly *definitional* effort: for all philosophers have typically cared, the theory begins and ends with a presentation of what it takes for one random variable (or the worldly feature it models) to carry information about another. But information theory goes well beyond that. It is, well, a *theory*, and as such it is chiefly composed of claims that are advanced in the hope that they be true about the world.

In a nutshell, the most celebrated results in information theory have to do with specifying how faithful the transmission of information from a source can be, when it happens over a (typically noisy, typically narrow) channel. These results have played absolutely no role in informational accounts of representation.[1] Take, for starters, the idealized depiction of an information-processing pipeline in fig. 1 (*cf.* Cover & Thomas 2006, fig. 7.1)

[1] Two recent philosophical treatments of information that try to redress this neglect are Mann (2018) and Rathkopf (2017).

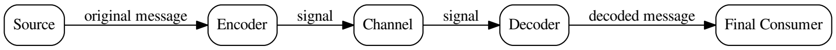

Figure 1: An information-processing pipeline

Here an *encoder* produces a signal as a response to information incoming from a source. This signal goes through a channel and is subsequently decoded, producing a message that is then utilized for whatever purposes downstream. The first thing to note is that the broadly Dretskean ideas about the content of a signal introduced in section 1 only have use for the first two links in this information-processing chain: how signals carry information about a certain original message produced by a source, as depicted in fig. 2. In fact, in information theory the main action happens immediately after that: a source is producing stuff, and we want that stuff to *go through a channel*. Information theory is mainly about providing theoretical guarantees of faithfulness in transmission, given the rate of the channel. We can think of this rate as the number of bits it provides for the encoder to use in the signal. If, say, the rate is 2 bits per use of the channel, this means the encoder can use up to 2 bits to construct the signal and be sure that it can pass unscathed through the channel and on to the decoder.

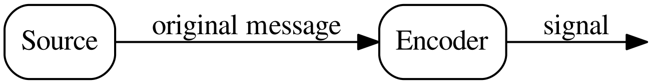

Figure 2: The information-processing pipeline in the Dretskean tradition

In typical cases of representation, channel rate is consistently smaller than ideal. Consider animal alarm calls. Vervet monkeys, for example, are typically described as being able to produce three different, discrete kinds of calls (Seyfarth, Cheney & Marler 1980a, 1980b) that are usually taken to be associated with the presence of leopards, eagles and snakes respectively. Obviously, the entropy of the relevant aspects of the environment that prompt the production of a call (think of all the possible patterns of approach of these predators, for example) vastly outstrip the rate of a channel, which consists in the production of just one out of three possible signals. This means that loss in communication is inevitable. Alarm calls, and for analogous reasons representations in general, are all about *lossy transmission*.

The way in which information theory deals with lossy transmission is by defining a *distortion measure* (Cover & Thomas 2006, p. 304) that gives a score to a pair composed of a certain original message M, and the decoded version thereof, $\hat{M}$. In what follows I will be using the *Hamming distortion* which simply adds 1 to the distortion when the bits in the original and decoded signals (which we can assume to be binary strings) do not coincide, and 0 otherwise, then normalizes. So, for example, the Hamming distortion between an original signal $M = 010011$ and a decoded signal $\hat{M} = 100010$ is $^3\!/_6$, because the first, second, and last (a total of 3) bits have been decoded incorrectly, and there are 6 bits in total.

The central result in this so-called *rate-distortion theory* approach to lossy transmission is that there is a *rate-distortion function*, $R(D)$, which gives the minimum rate at which any given distortion is achievable. The actual mathematical expression of the rate-distortion function need not detain us here (see Cover & Thomas 2006, p. 307, theorem 10.2.1), but it is such that the *Blahut-Arimoto* algorithm (Blahut 1972; Arimoto 1972) allows us to calculate it easily.

The main thesis of this paper is that representations belong in information-processing pipelines whose rate-distortion function has *sweet spots*: by this I mean points in the rate-distortion curve such that the usefulness of increasing the rate of the channel past those points is much smaller than before reaching them. Moreover, the encoding-decoding strategies that make use of these representations tend to live in the vicinity of those sweet spots. I submit that it is these information-theoretic properties that the conditions on representation discussed in section 2 try to get at.

To see how rate-distortion analyses work let's start by looking into a source that models a series of fair-coin tosses: this random variable would have two values, *heads* and *tails*, with associated probabilities $P_{heads} = P_{tails} = .5$). Using the Hamming distortion as our target distortion measure, if the coin lands heads (tails) and the decoded message is tails (heads) the distortion is 1, otherwise 0. The Blahut-Arimoto algorithm allows us to draw the rate-disortion curve, in fig. 3. Here the blue line is the rate-distortion curve. It intersects the x-axis at 1.0 bits (the entropy of the source) and it intersects the y-axis at 0.5 (the lowest average distortion one can achieve when the channel is closed.) The red line gives a measure of how steep the blue line is at any given point—in particular, the absolute value of the slope of the blue line. The higher the red line, the steeper the blue line.

The situation this setup is modeling is one in which a single cue is present or absent, and a signal tries to keep track of whether it does. This is precisely the kind of situation where many theorists (certainly Sterelny and Burge, for the reasons reviewed in 2.1) would see the postulation of representations as entirely idle—see, e.g., Schulte's vasopressin example in his Schulte (2015). In agreement with the idea that postulating representations here is idle, there is not much structure to the rate-distortion curve corresponding to this setup: reading the chart from right to left, increasing the rate makes the achievable expected distortion go smoothly down, until the rate hits the entropy of the source, at which point the achievable distortion is zero.That's about it.

Let's now model one kind of situation in which there is a reasonably wide consensus that representations make an explanatory contribution: vervet-monkey alarm calls, as reviewed above. In the model, the source—the situation the information-processing pipeline is dealing with—randomly makes members of two natural kinds (we can think

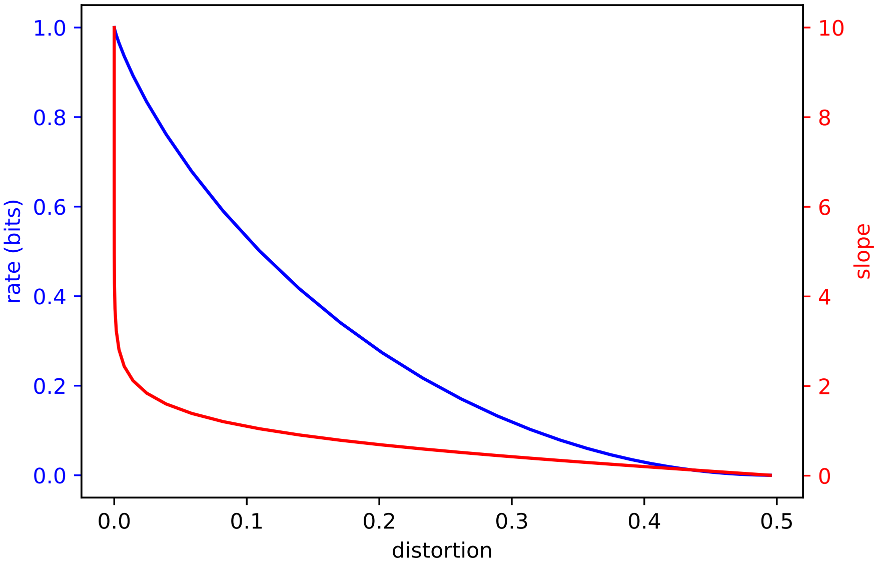

Figure 3: The rate-distortion function for a coin toss

of them as two different predators) be or not present at any given time, independently from one another. This intends to mimic the situation vervet monkeys face, where snakes, leopards and eagles show up or not, more or less at random.

These natural kinds are modeled as homeostatic property clusters (see section 2.2 above). In order to derive a explicit probability distribution for the source out of this qualitative description, the two HPCs are in their turn represented by two Bayesian networks, each with a parent node and four children (see fig. 4.) Each of the nodes stands for a property; if the node is *on* it means the corresponding property is instantiated; if it is *off* it means it is not. In the model, children nodes replicate noisily the state of their parent. Thus, e.g., if the parent is *on* (if the corresponding property is instantiated) each child property will have a .95 chance of being instantiated too; if the parent is *off* the probability for each children of being instantiated is .05. The unconditional probability of instantiation for the two parent nodes is .5.

In the model, the source produces a binary string, with each member of the string being 1 if the corresponding node is on, and 0 if it's off. This signal is encoded, goes through a channel, and is then decoded at the other side. The target distortion measure is the Hamming distortion. Fig. 5 plots the rate-distortion curve for this model.

This curve is very different from the one in fig. 3: there is a clear "sweet spot"—a sudden drop in the usefulness of extra rate, see the red curve—when the system hits a rate of 2 bit/use. I.e, there is, in a certain principled sense, an optimal level of lossy compression; a way to set up an encoding-decoding strategy that recover most of what's going on in

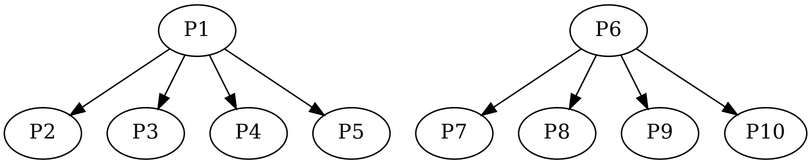

Figure 4: Two natural kinds

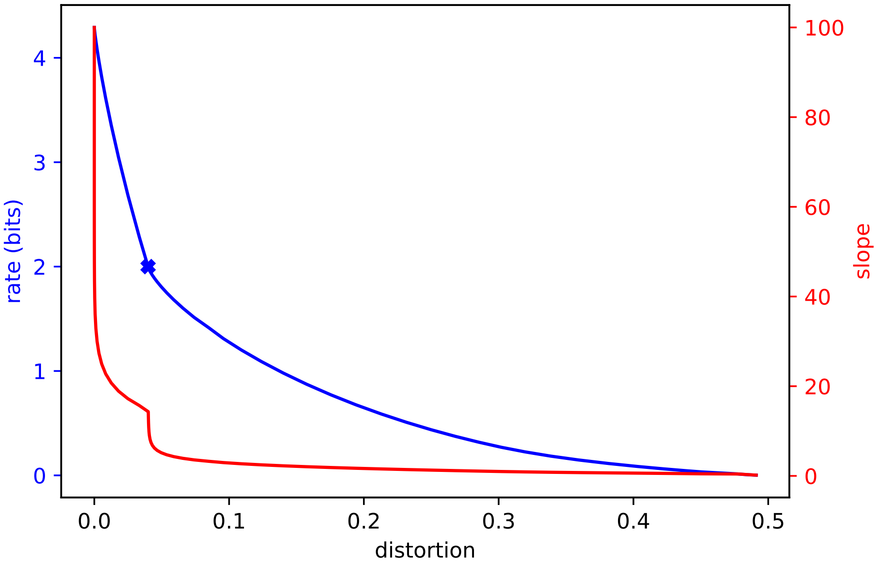

Figure 5: A sweet spot in the rate-distortion function

the world of relevance to the information-processing system, even through a very severe, 2 bit bottleneck. I claim that this is no coincidence. Our representation-attributing practices gravitate towards this kind of situations.

To see how sweet spots in rate-distortion curves and representations are related, consider now what an optimal encoding-decoding strategy would look like. That is, how should the encoder encode the information coming from the source, and how should the decoder decode the signal coming from the encoder, so that the resulting expected distortion between original and decoded signal is the minimum achievable, at the sweet spot?

Optimal Encoding Strategy: First divide the incoming signal in two halves, one corresponding to properties P_1 through P_5; the other corresponding to properties P_6 through P_{10}.

If there is a majority of 1s in the first half of the original signal set the first bit of the signal to 1. Otherwise set it to 0. Ditto for the second half of the original signal and the second bit of the signal.

Optimal Decoding Strategy: If the first bit in the incoming signal is 1, set the first half of the decoded signal to 11111. Otherwise, set it to 00000. Ditto for the second bit and the second half of the decoded signal.

How should we interpret what encoder and decoder are doing here? A natural way is this: they are using the presence or absence of properties in an HPC cluster as diagnostic of the presence or absence of the underlying natural kind—this would be the encoding part—and then taking the resulting signals as representing the presence of a paradigmatic instance of the kind, one that has all the properties in the cluster—this would be the decoding part. HPC kinds being what they are, frequently the first half of the incoming signal will resemble the paradigmatic presence of the first kind (11111) or its paradigmatic absence (00000), and the same will happen with the second half and the second kind. That is why this encoding-decoding strategy works so well.

In describing this optimal strategy I have helped myself to representational vocabulary; it has been useful in order to explain how the strategy works, and how come that behaving in this particular way achieves low distortion at low rates: it is because each of the two bits in the signal is caused by, and causes, behavior that is optimally attuned to the probabilistic structure of each of the two natural kinds in the model world, respectively. Nothing going on in this system falls outside the purview of Shannonian information theory—of information theory *tout court*, so at least in this kind of cases representational talk depends on no non-informational fact.

We can now understand better what's lacking in the philosopher of mind's information-theoretic tookit: it is entirely possible, and computationally trivial, to calculate, e.g., Skyrms's pmi between each of the possible signals (00, 01, 10 and 11) and each of the possible world states (all 1024 of them, from 0000000000 to 1111111111). Doing so would leave us with 4 vectors (one for each signal) with 1024 entries each (one for each world state.) First, this is an unwieldy collection of numbers, which doesn't bring out the relevant structure. For example, if the probability of children nodes being *on* conditional on their parent being *on* was .96 instead of .95 the rate-distortion curve would be qualitatively identical, with a sweet spot in exactly the same place, yet most numbers in the Skyrmsian informational content vectors would change. Second, and most important, nothing in those 4096 numbers allows us to infer the presence of a sweet spot. The relevant information is simply not there, depending as it does on a distortion measure which is not used in computing Skyrmsian informational contents.

If this is approximately right, the question about what makes representational talk explanatory is readily answered: saying that a certain vehicle is a representation conveys something quite specific about its informational context. It says that the vehicle is part of an encoding-decoding strategy that exploits a sweet spot in a rate-distortion curve—where the curve is in turn fixed by the probabilistic structure of the world, and the target distortion measure. This, in less technical terms, translates to saying that the vehicle is summarizing *relevant* (this is where the distortion measure comes in) aspects of the current situation in an optimal, if lossy, manner, made possible by *how the world* is (this is where the probabilistic structure of the world comes in.) This explication of the explanatory contribution of representations can be turned into an explicit answer to what makes something a representation—an answer, that is, to what Artiga (2016) calls the metasemantic question.

The Rate-Distortion Approach: A signal, S, in a certain information-processing pipeline, P, is a representation if the following two conditions are met:

Existence: There are sweet spots in the rate-distortion curve associated with P.

Optimality: S is produced as part of an encoder-decoder strategy that occupies the vicinity of one of these sweet spots.

So, *pace* Dretske, the core information-theoretic notions of entropy, rate, distortion, etc. *can* provide invaluable insight into the representational status of individual signals. If the rate-distortion approach is on the right track, those information-theoretic notions, through the existence condition, specify the kind of setup where representations live, which then the optimality condition can use to provide a criterion for the representational status of individual signals.

I offer the foregoing discussion as a preliminary case for the rate-distortion approach to representation: it shows how postulating representations is explanatory, even if these representations depend just on (Shannon) information. It illuminates the difference in representational status between cue-driven examples, such as Schulte's vasopressin; and vervet alarm calls, and other similar examples. To complete my case I now show how the ways to bridge the gap between natural and nonnatural information discussed in section 2 can be seen as unwitting attempts to get at rate-distortion sweet spots.

4 There is no Gap to Bridge

What does it take for the existence condition to be met? That is to say, what circumstances result in sudden drops in the slope of the rate-distortion curve? We have seen one such family of circumstances: if the pattern in which properties are instantiated in the source is noisily replicated in a cluster then sudden drops are to be expected: distortion will decrease with rate up to the point where all the main sources of variation in property instantiations are accounted for, and all that remains is the residual noise in instantiations within each cluster. Take a look again at figs. 4 and 5: to describe this source we basically need enough rate to account for the two main sources of variation: P_1 and P_6. This is not all there is to the world, because it's possible for the other properties to (fail to) token independently of their parent, but the unlikeliness of these departures makes the extra rate comparatively less useful.

Noisy replication of property instantiations is at the core of the HPC theory of natural kinds, as we saw above. This means that, in general, the presence of HPC natural kinds in a source will create sweet spots. This opens a line of argument in favor of reference magnetism from information-theoretic premises: reference magneticsm should be seen as making a point about the kind of probabilistic structure that an information-processing pipeline must be attuned to, if signals are to effect the kind of optimal lossy compression that underlies our representation-attributing practices. Reference magnetism is just a way of meeting part of the existence condition.

Regarding the suggestion, by Sterelny, Burge and others, that representations inhere preferably on signals sitting in a one-to-many-to-one pipeline, I submit that the many-to-one aspect of this suggestion aims at meeting the optimality condition; the one-to-many aspect, together with reference magnetism, aims at meeting the existence condition.

The first thing to note here is that the *Optimal Encoding Strategy* presented above enforces what Sterelny calls robust tracking and Burge calls constancy: the strategy consists in considering all properties coming from each of the two clusters and setting the relevant bit to 1 only if a majority of those properties are instantiated. That is, the encoder is taking a multiplicity of configurations (e.g., the first half of the incoming signal being 00111, 01011, 10111, etc.) to a single output: the first bit of the signal being 1. Furthermore, that part of the signal will be decoded as 11111: from there on, the system downstream will treat whatever is out there in the world as a paradigmatic member of the first kind. The system is recovering the presence of a natural kind out of many different, noisy instantiation patterns. This is a clear instance of constancy. Suppose that the encoder, insted of being many-to-one, depended on a single cue; say, suppose it set the first bit to 1 if one of the children properties (say, P_2) was instantiated, and to 0 otherwise. In such a cue-driven setup, the best encoder-decoder arrangement possible is marked by the blue circle in fig. 6. This has double the distortion than than the optimal encoding (marked by the blue cross) which sits right on top of the optimal rate-distortion curve. This cue-driven system would not meet the optimality condition, which means that a many-to-one architecture is instrumental to meeting it.

Finally, the target distortion measure in the information-processing pipeline can be seen as that which Sterelny's one-to-many condition on representation is actually tracking. Using, for example, the Hamming distance as a distortion measure is tantamount to assuming that all of the properties of the natural kinds are relevant for downstream processing. One natural way in which this may happen is when the agent is to respond flexibly to the presence of the natural kind: in different contexts or states different properties of the kind might be relevant and, for example, the presence of a tree might

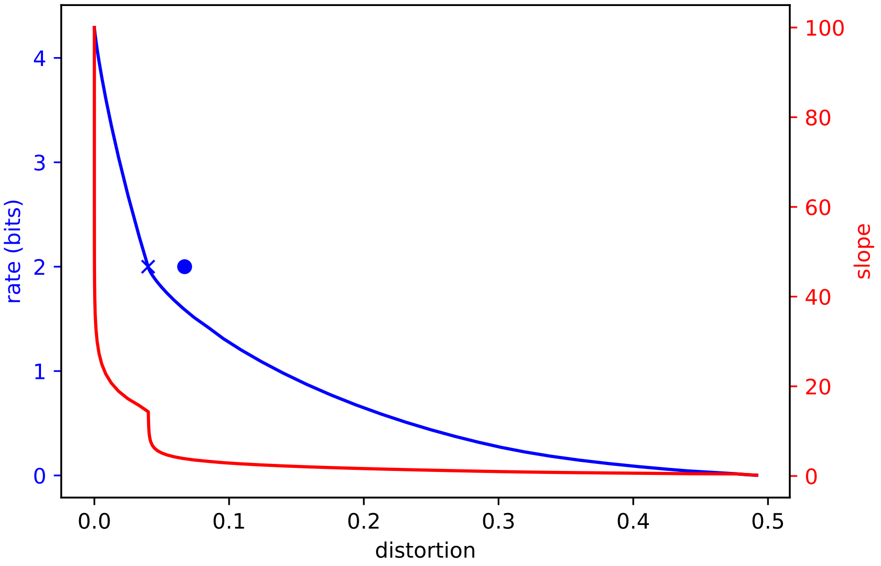

Figure 6: Cue-driven encoding

be sometimes relevant to behavior because it bears fruit (if the agent is hungry) and some other times because it has a dense cover (if the agent is looking for shelter.)

Caring about all (or many) properties of the kind is what makes the rate-distortion curve display a sweet spot. If, instead, the agent has a rigid, stereotyped response to the presence of members of the kinds—that is, if it only cares about the presence of one property, which is the property that makes that rigid behavioral response fitness-conducive, then the curve is as presented in fig. 7. Rigid behavioral responses make the probabilistic structure of the kinds largely irrelevant. As a result, the system behaves as if a coin were tossed, where heads would mean that the target property is tokened, and tails that it is not. This arrangement does not meet the existence condition. Sterelny's broad-banded responses are, again, a way of getting at rate-distortion sweet spots.

References

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Artiga, M forthcoming, 'Beyond Black Spots and Nutritious Things: A Solution to the Indeterminacy Problem', *Dialectica*.

Artiga, M 2016, 'Liberal Representationalism: A Deflationist Defense', *dialectica*, vol.

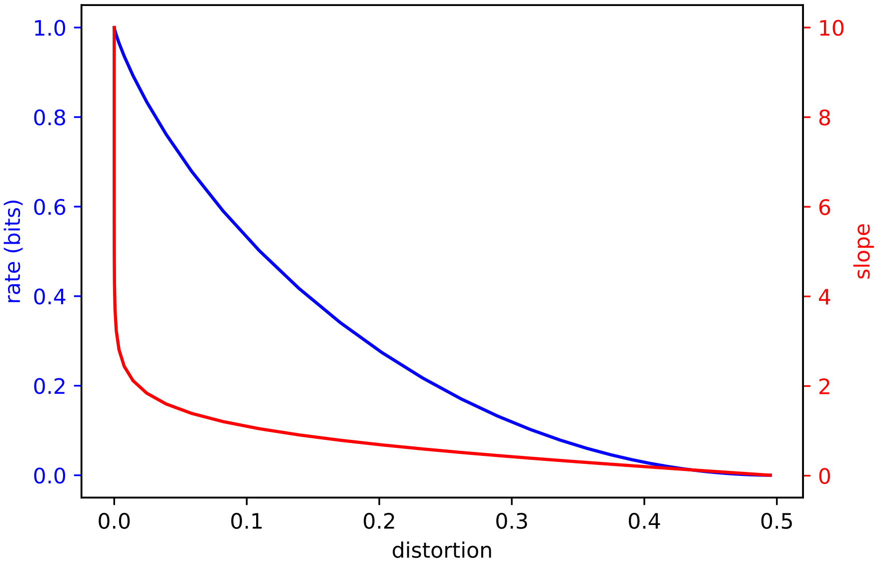

Figure 7: Rigid behavioral response

70, no. 3, pp. 407–430.

Blahut, R 1972, ‘Computation of channel capacity and rate-distortion functions’, *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Boyd, R 1989, ‘What Realism Implies and What It Does Not’, *Dialectica*, vol. 43, no. 1-2, pp. 5–29.

Burge, T 2010, *Origins of objectivity*, Oxford University Press.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

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

Lewis, D 1983, ‘New work for a theory of universals’, *Australasian journal of Philosophy*, vol. 61, no. 4, pp. 343–377.

Lewis, D 1984, ‘Putnam’s paradox’, *Australasian Journal of Philosophy*, vol. 62, no. 3, pp. 221–236.

Mann, SF 2018, ‘Consequences of a Functional Account of Information’, *Review of Philosophy and Psychology*, pp. 1–19.

Martínez, M 2013, ‘Teleosemantics and Indeterminacy’, *Dialectica*, vol. 67, no. 4, pp.

427–453.

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

Papineau, D 1987, *Reality and Representation*, Basil Blackwell.

Rathkopf, C 2017, ‘Neural information and the problem of objectivity’, *Biology & Philosophy*, vol. 32, no. 3, pp. 321–336.

Ryder, D 2006, ‘On Thinking of Kinds’, in G Macdonald & D Papineau (eds), *Teleosemantics*, Oxford University Press, pp. 1–22.

Ryder, D 2004, ‘SINBAD Neurosemantics: A Theory of Mental Representation’, *Mind & Language*, vol. 19, no. 2, pp. 211–240.

Scarantino, A 2015, ‘Information as a probabilistic difference maker’, *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Schulte, P 2015, ‘Perceptual representations: A teleosemantic answer to the breadth-of-application problem’, *Biology & Philosophy*, vol. 30, no. 1, pp. 119–136.

Seyfarth, RM, Cheney, DL & Marler, P 1980a, ‘Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication’, *Science*, vol. 210, no. 4471, pp. 801–803.

Seyfarth, RM, Cheney, DL & Marler, P 1980b, ‘Vervet monkey alarm calls: Semantic communication in a free-ranging primate’, *Animal Behaviour*, vol. 28, no. 4, pp. 1070–1094.

Sider, T 2014, *Writing the Book of the World*, Reprint edition., Oxford University Press, Oxford.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Sterelny, K 2003, *Thought In A Hostile World: The Evolution of Human Cognition*, John Wiley & Sons, Malden, MA.