PhiMiSci
Philosophy and the Mind Sciences

The information-processing perspective on representation

Manolo Martínez[a] (mail@manolomartinez.net)

Abstract
I introduce a novel framework for theorizing about representations in cognitive science, which relies on two theses. First, representations are, primarily, signals for information transmission, not as a side effect of other functions these signals may have, but for its own sake. Second, these signals aim at efficiently trading-off three cognitive budgets: rate (or transmission and storage costs), distortion (or faithfulness of the transmitted information), and computational complexity of coders. The way to provide empirical evidence that some entity is a representation—and hence that it is involved in information transmission for the sake of information transmission, as per the first thesis—is to show that it present adaptations for efficient information transmission—as per the second thesis. The kinds of properties that cognitive scientists routinely associate with paradigmatic instances of representations are generated by adaptations for rate-distortion-complexity efficiency.

Keywords
Complexity · Distortion · Information · Rate · Representation

1 Introduction

Here is a relatively commonplace observation: representations, to the extent that they can be found in cognitive systems, are involved in the transmission of information—notably, but not only, from the sensory periphery to the production of behavior. My objective in this piece is to sharpen this observation into a substantive theory of representation. More concretely, I will argue for two theses:

Transmission: Representations are, primarily, vehicles of information transmission for its own sake.

RDC Trade-Off: Representations aim at efficiently trading-off three key cognitive budgets:

1. *Rate* of signals (transmission and storage costs),
2. *Distortion* of signals (faithfulness of the transmitted information), and
3. Computational *Complexity* of coders.

[a] Universitat de Barcelona.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321
 https://philosophymindscience.org ISSN: 2699-0369

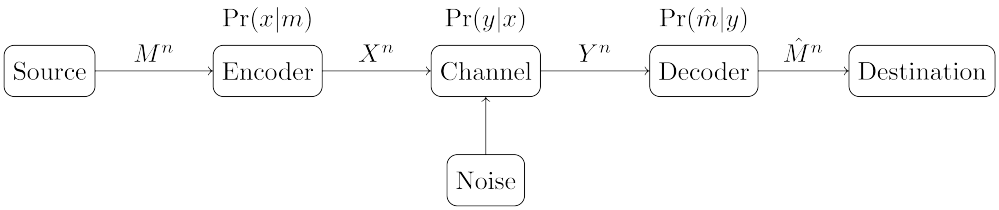

Figure 1: A point-to-point information-transmission model (redrawn with some modifications from Shannon, 1948)

This is how these two theses represent a sharpening of the above observation. The Transmission thesis claims that information transmission is not just something that representations happen to do, but what they are *for*. To theorists educated in an understanding of information as a quantification of dependences among variables (e.g., Dretske, 1981; Neander, 2017; Skyrms, 2010) this will sound like the familiar claim (Dretske, 1988, Chapter 3) that representations are entities meant to stand in relations of correlation or dependence with other, often extramental, entities. This is not what I mean. I mean the strictly stronger claim that representations are signals that present adaptations (compression, error protection, complexity management, see below) that further the goal of transmitting information *per se*, to a significant extent independently from whatever that information is about—adaptations for information transmission *for its own sake*, as I will sometimes say.

One general model of information transmission for its own sake is provided by information theory.[1] In §3 and §4 I say more about what this model entails, but for now we can summarize it as one in which one aims to get information from a source where it is generated, to a destination where it is put to use, while negotiating several constraints on this transmission. This is done with the help of a family of coders, which prepare the signals that carry that information so that they can best withstand those constraints, and subsequently convert them again in whatever ways are most conducive to the downstream uses the information is to be put to; see Figs. 1 and 2.[2] The fact that the destination variable is tailored to downstream uses means that it may possibly, and in fact typically, look very different from the source variable. That is to say: importantly, *information transmission, in the general sense I am outlining here, also includes information processing and transformation.*

Very often in cognitive science and its philosophy, appeals to information are restricted to quantitative versions of the locution "A carries information about B", where, typically, A is a signal and B a worldly variable (Neander, 2017; Shea,

[1] By "information theory" I mean the theory of information pioneered during the mid-XXth century heyday of cybernetics by Claude Shannon, Norbert Wiener and others. Shannon (1948) is the foundational text.

[2] In this piece I use *coder* as an umbrella term for both encoders and decoders, and *codec* for an encoder-decoder pair.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

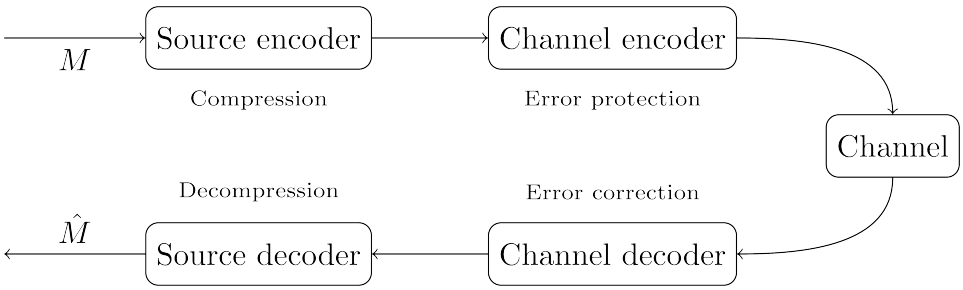

Figure 2: A closer look at codecs

2018, sec. 4.1). This way of thinking of information encourages a view of signals as entities that systems encounter and exploit as a means to know about the world upstream—modeled after how, e.g., smoke is encountered and exploited as a means to know about fire. In information-theoretic models, instead, information is thought of as flowing from source to destination, carried by signals that are *designed* or *evolved* (not merely encountered) to deal effectively with transmission constraints. In some respects this is just a difference of emphasis, but it is an important one: as I will show, constructing signals that transmit information efficiently is a difficult engineering problem, not just something one stumbles upon. If and when evolved systems solve this problem, this provides evidence that transmission of information for its own sake was the goal.

This suggests a way of gathering empirical evidence that a certain state or process in an evolved system has an information-transmitting role—not just that it happens to transmit information in the Dretskean, correlational sense, but that information tranmission is one of the things it is tasked with doing—and therefore (as per the Transmission thesis) that it is a representation: we look for evidence that it is solving the difficult engineering problem that is efficient information transmission. This is where the RDC Trade-Off thesis comes in. Information transmission is all about getting task-relevant information (i.e. relevant to whatever task the cognitive system is engaged in, see §2) from source to destination, while negotiating various constraints. One natural suggestion as to ways to anchor the Transmission thesis empirically is, first, to identify a set of constraints that information transmission is specially sensitive to, and then look for adaptations aimed at negotiating these constraints in particular. The RDC Trade-Off thesis makes the concrete bet that there are three main families of constraints on information transmission, corresponding to three main families of "cognitive budgets"—distortion, rate, and complexity—that need to be jointly minimized:

Distortion: Information-transmission systems aim at sending as much *relevant* information as possible from source to destination.

Equivalently, the Distortion constraint enjoins information transmission to be as (relevantly) faithful as possible. More about distortion measures in §2.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

Rate: Information-transmission systems aim at sending as little information (*tout court*) as possible from source to destination.

In a nutshell, trading off rate and distortion amounts to being as communicatively circumspect as possible, while still managing to make the destination variable sensitive enough to the source variable for the current goals of the system.

While characterizing and quantifying rate-distortion trade-offs is one of the main goals of information theory, characterizing and quantifying our third cognitive budget is not. In information theory, coders are thought of as black boxes that effect the requisite signal transformations. The focus is on how much coders, any coders, can compress signals, or make them withstand noise; but their internal makeup, the means by which they should do this, is essentially glossed over. It is assumed that, when a value of the source variable is fed to the encoder, the desired signal comes out; and that, when a signal is fed into the decoder, the desired value of the destination variable comes out. Of course, carrying out these transformations from variables to signals and back is, in the general case, far from trivial, and different ways to do it incur in different costs. The third budget comes in at this point.

Complexity: Information-transmission systems aim at keeping codecs as computationally simple as possible.

Rate and Complexity constraints can be seen as enjoining the information-transmission system to be as economical as possible. Rate is mostly about being *dynamically* economical: in the production of destination variables from source variables, as little as possible should be transmitted. Complexity is mostly about being *statically* economical: the resources used to construct signals out of source variables and to construct destination variables out of signals should be as lean as possible.

The picture that emerges from the Transmission and RDC Trade-Off theses is one in which cognitive tasks conjure up an abstract space of possible solutions, with rate, distortion, and complexity dimensions. Within this space, there is a *Pareto frontier*: a set of RDC trade-offs that are optimal in the sense that we cannot decrease the amount of any of the budgets that is used without simultaneously increasing the amount of at least one of the other two budgets. This Pareto frontier I will call the *representational surface*.

I offer the combination of Transmission and RDC Trade-Off theses, together with the attendant notion of a representational surface, as a useful framework in which to think of representations in cognitive science. As I show in §4 and §5, cognitive scientists *already* routinely, if often tacitly or even unwittingly, rely on analogues of these theses in their attribution of representational status to cognitive states and processes. This is how it works: in information theory, and to a lesser extent complexity theory, RDC budgets come together with attendant operations aimed at minimizing them. *(Shannon) compression* and *error correction* are

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

the operations aimed at optimizing rate-distortion trade-offs; while what we could call *(algorithmic) compression* is one operation, among others, aimed at optimizing complexity-distortion trade-offs. In §4 and §5 I show that cognitive scientists very often appeal to the presence of informal analogues of these operations in the characterization of paradigmatic examples of representational systems: what counts as a good example of a representation often is an entity that is recognizably RDC-efficient. This is not a coincidence, of course—the present framework was developed by paying attention to, and in order to accommodate, uses of representational notions in cognitive science. In §7 I offer some concluding remarks, and some pointers for further research.

One final word about the aims and scope of this article. This is a position paper: I aim at identifying, clarifying and staking out a certain, so far underdeveloped, position in the debate over representation in cognitive science—the one I have summarized in this introduction. As such, the discussion will be comparatively light on the kind of polemical material (preemptive responses to objections, critiques of other accounts, etc.) that figure prominently in much philosophy of cognitive science. This kind of material is, of course, necessary to make a full case in favor of the theses presented here. Some of it has appeared in (De Llanza Varona & Martínez, 2024; Martínez, 2019b, 2024), and more will follow. Here, though, I focus on providing as clear an exposition of the main idea as possible.

2 Information and adaptive behavior

How is information transmission important to cognition? Cognition appears to be closely linked to the production of "adaptive behavior" (Barack & Krakauer, 2021, p. 359)—that is to say, behavior that is attuned to "the structure of the environment and the goals of the [cognitive agent]" (Anderson, 1990, p. 3). The consensus around this idea extends well beyond classical representationalist cognitive science, to theorists who favor non-representationalist accounts of cognition: e.g., Varela et al. (2016, p. 173) claims that perception is best thought of as "perceptually guided action" (see also Noë, 2004); according to Brooks (1991, p. 145) an intelligent being "must cope appropriately and in a timely fashion with changes in its dynamic environment"; for Chemero (2011, p. 212), cognition involves "closely coordinated perception and action". Theorists interested specifically in the demarcation of cognitive and non-cognitive phenomena make similar claims. Akagi (2021, p. 7), for example, argues that cognition is "the sensitive management of an agent's behavior."

If cognition is at least partly, at least some of the times, about producing *behavior* that is *appropriate* to the *current circumstances* of the agent, then information transmission is necessary for cognition.[3] We can clarify why this is so by focusing

[3] I do not claim that it is sufficient for cognition, though. As I will explain, considerations of efficiency regulate our attributions of representational status, and it is entirely possible that some

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

on the three concepts in italics in the previous sentence. First, we can model the subject's current circumstances as a *random variable*, C: that is to say, as a collection of possible values for the variable to take (i.e., circumstances for the agent to be in), together with a probability distribution over those values, $\Pr(C)$.[4] For concreteness, think of a sensorimotor-control task, such as grasping (e.g., Smeets & Brenner, 1999) or pointing (e.g., Körding & Wolpert, 2004). For any of these tasks, circumstances will consist, minimally, of the position, shape, and size of one or various objects in the vicinity of the agent, with the different possible values of C corresponding to the different possible ways these properties can be specified. In a laboratory task, it is the researcher who fixes the frequencies with which each of those values (each type of stimulus) will be presented to the subject. In more naturalistic settings, the world takes care of that. The same kind of thinking can be applied to behavior: this, too, can be modeled as a random variable, B, in which values correspond (minimally, and glossing over important philosophical complications, e.g., Millikan, 1993; Shepherd, 2021, Chapter 2) to specifications of patterns of muscle activation and joint torques, among other things (Todorov, 2004).

We have now two variables, C and B. As per the insight rehearsed above, one of the main goals of cognition is producing values of B (this production is, typically, more or less under the control of the agent), *appropriately*, in response to values of C (which are typically not under the control of the agent.) The third and final component in this barebones operationalization of cognition has to do with what counts as appropriate. The modeling decision that is universally taken at this juncture is to assume the existence of a function giving a score (say, a positive real number) to each pair of a value of C and a value of B:

$$d : C \times B \to \mathbb{R}^{+}$$

This function d, a *distortion measure* (Berger, 2003), crops up everywhere, under different names and very slightly different guises, wherever attempts are made to formalize cognitive tasks. In machine learning and computational neuroscience it is called the *loss*, *cost*, or *objective function* (Chollet, 2021, p. 9; Dayan & Abbott, 2001, Chapter 3); in decision and game theory it is called the *utility function*, or simply the *payoff table* or *matrix* (Myerson, 1997, p. 5; O'Connor, 2020, p. 6; see also Martínez, 2019a for more on the relation between these superficially different measures). In this paper I will use "distortion measure" as my preferred name for d, with the understanding that lower distortion is better, but this choice is immaterial,

instances of information transmission (at least in the liberal understanding of this notion I am relying on here) do not qualify.

[4] We often need to consider sets of various random variables, M_i, the probabilities associated to which are given jointly, $\Pr(M_1, \ldots, M_n)$. Also, typically, cognitive scientists do not have access to variables so defined but to series of experimental values. Random variables can still be constructed in these cases, by binning and calculating empirical measures over those series (Timme & Lapish, 2018, p. 7).

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

and I invite the reader to substitute it by whatever other terminology ("objective function", "loss" "faithfulness", "fidelity", or what have you) they feel more comfortable with. A more formally perspicuous specification of a central kind of cognitive function is, then:

Main Model: The transformation of one variable, C, into another, B, in a way that minimizes[5] a certain distortion measure $d(C, B)$.

I claim that Main Model captures the standard way in which cognitive tasks are operationalized in much of cognitive science, including computational neuroscience and machine learning. Now, adequately performing cognitive tasks that fall under the Main Model requires information transmission. This is the intuitive picture: the distortion measure d enforces a preference for certain values of the output variable, B, given the current value of the input variable, C. This is simply what d is. This, in turn, means that, if the system is to perform the task in question adequately (that is to say, if it is going to achieve a low value of d, or at any rate a value lower than random behavior could achieve) it needs to select reasonably good values of B for any given C: movements of the hand that get it closer to a cup, if the task is to pick up the cup. This is the essence of information transmission: the current value of C (e.g., position and shape of the cup) constrains the probability of current values of B. In the next section, I will be more explicit about how this dependence can be expressed, but the idea is that *adequately performing a Main-Model task requires information transmission in the sense of task-relevant reduction of uncertainty*. Reduction of uncertainty in that the probability of different values of B is constrained by the value of C. Task-relevant in that it only cares about information that directly helps reduce the value of d.

Importantly, information flow from one variable to another is typically impeded by various kinds of constraints. As we will see, it is the negotiation of these constraints that generates representation-like properties. I start to develop this idea, in the next section, by explaining how the foregoing notions are formalized in information theory.

3 Informational constraints on transmission

In the philosophy of cognitive science, information theory is often thought of as a set of measures of correlation and dependence between variables (Isaac, 2019; Scarantino, 2015; Sprevak, 2019). In cognitive science and neuroscience, too, the stress is often put on how information theory quantifies "how much information a neural response carries about the stimulus" (Borst & Theunissen, 1999, p. 947; see also Rieke et al., 1999, p. 102), and, in general, on how it deals with "statistical aspects of [the] correspondence [between two domains]" (Brette, 2019, p. 1). But

[5] "Minimizes" just means here that smaller is better. Nothing relevant changes if you prefer to think of this in terms of satisficing.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

the main goal of information theory is not to provide alternatives to other well known measures of correlation and dependence among variables—although that is something the theory does. It is, rather, to provide a firmer grip on the elusive notion that information can be generated at some place or time and transmitted to another; and, more importantly, to offer ways to calculate optimal bounds for this transmission.

The way information theory tackles its first goal is well known, and I will not go through the details here (they can be found in any of a number of standard textbooks, such as Cover & Thomas, 2006; or MacKay, 2003): the quantity of information in a random variable M, also called its *entropy*, $H(M)$, in bits, can be thought of as the average number of yes/no questions one needs to ask in order to know the current value of M. A variable has nonzero entropy if it can take more than one possible value (and therefore we need to ask at least one yes/no question sometimes, in order to know which value it is). This is what it means for a random variable to *produce* information: its taking different values on different occasions, in a way governed by the probabilities of its values and summarized by its entropy. And that is why we call the first link in the well known "point-to-point" information-transmission model (Fig. 1) a *source* of information: it is a source of uncertainty—a random variable we (or someone, or something) would be better off knowing the value of. We say that information flows from one variable M to another variable $\hat{M}$ to the extent that knowing the value of $\hat{M}$ reduces the system's uncertainty about M. Calculating this reduction in uncertainty, the so called *mutual information* of the two variables, $I(M;\hat{M})$, just involves calculating the expected value of the entropy of M, using the probability of M values conditional on the different possible values of $\hat{M}$, instead of their unconditional probability.

Mutual information and related measures are sometimes used in cognitive science as evidence in the reconstruction of mechanisms (Borst & Theunissen, 1999; Wei & Stocker, 2015): if information has flowed from a worldly variable M to a neural variable X, in the sense that the mutual information between these variables is nonzero, this provides evidence that X participates in the processing of information incoming from M (Fig. 3). Just fallible evidence, of course: if we have no further evidence of downstream use of the information, it may be that this information about M is there simply as a byproduct of other processes with other goals (cf. Quian Quiroga & Panzeri, 2009, p. 178; Rathkopf, 2017; Wachtler et al., 2003, p. 689).

This is precisely why Shannon's point-to-point model of communication (in Fig. 1), while very simple, is more complex than a mere recording of dependences among variables (as in Fig. 3): the aim of information theory is not merely to measure probabilistic dependence, but to estimate bounds to the faithful transmission of information in the presence of various constraints. This will of course involve information flowing from an input variable (the *source*) to an output variable (the *destination*), but focusing on constraints to transmission, and on the way to overcome them, means that we need to peer into the process that takes information

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

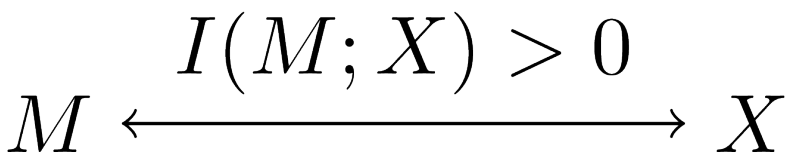

Figure 3: Simple informational dependence between two variables

from source to destination, and make modeling assumptions about what this process will entail. These assumptions are represented in Fig. 1, which can therefore be seen as spelling out what, in §2, I called the Main Model of cognition.

1. A source (often some aspect of the external world that is out of the control of the cognitive system in question) emits a message out of a possible menu of options M.

2. Information about this message needs to be squeezed through a *channel* that models the difficulties that passage of information needs to face—a channel simply consisting in a collection of probabilities of sending in one signal and getting a different one at the other side. See the third link in Fig. 1. Information theory summarizes these probabilities as the *capacity* of the channel: the maximum possible mutual information between its input and output variables, X and Y (Cover & Thomas, 2006, p. 184).

3. Information about M is used in the construction of a destination variable, $\hat{M}$, say, a behavioral one.

With this description in hand, we can finally start looking into the first two "cognitive budgets" in the RDC Trade-Off thesis above—Rate and Distortion; and the operations by which they are optimized.[6]

4 Rate-distortion trade-offs

I can now be more precise about the claim made in §2, that reducing distortion in a Main-Model task necessitates the transmission of information. Shannon's lossy source coding theorem (1998–1949, theorem 21), establishes that *the only way to*

[6] It is common for philosophers, at least since Bar-Hillel & Carnap (1953), to claim that information theory in the Shannonian tradition is irrelevant to semantics. Mann (2023) does an excellent job of untangling some of the misunderstandings on which this claim rests.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

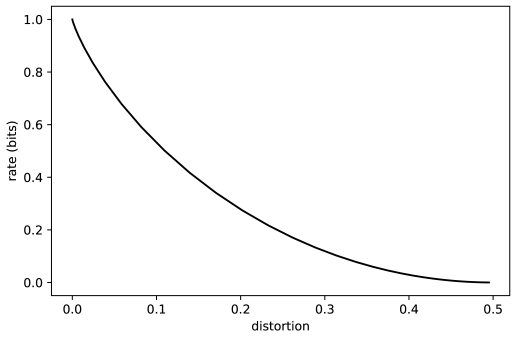

Figure 4: A typical example of a rate-distortion function. Here the source is a fair-coin toss (two possible, equiprobable values), and the distortion measure is the Hamming distance: 0 if original and destination messages are equal, 1 if they are not.

lower expected distortion is by expending more information in communicating the value of the source variable. The idea is the following: suppose that your source is spitting out values, and you want to send information about those values, so that the destination can produce the best possible match. In order to achieve this, you need to construct a signal that recovers whatever in the source variable is most relevant to the destination variable, and send it their way. This is what the *source encoder* does—see Fig. 2. When I say that we can use more or less information to communicate the value of the source variable, what I mean is that the encoder can use more or less *rate*: the expected number of bits the encoder can use to construct the signal corresponding to each source value—i.e., the richness of the repertoire of signals it can produce.

More precisely, the lossy source coding theorem states that for each possible expected distortion there is a minimum rate that achieves it; and vice versa: for each possible rate there is a minimum expected distortion that a encoder can achieve. There is a one-to-one correspondence, $R(D)$, between expected distortions and rates—an optimal compression rate for any desired level of distortion. Fig. 4 is a typical example of a rate-distortion curve.

Meeting the $R(D)$ optimum, for any given rate, is not trivial: if you want to get as much reduction in distortion from your available rate as possible you need to use sophisticated coding techniques (Sayood, 2017). The crucial point for my current purposes is that, while information merely flowing from source to destination might just be a byproduct of the destination being probabilistically sensitive to the source (because there will be nonzero mutual information between any two variables that are not probabistically independent from one another); *this information flow happening near a rate-distortion optimum*, on the other hand, is excellent evidence of design or adaptation for the transmission of information for its own sake: the causal connection between source and destination has been reshaped in a way that makes perfect sense if there is a goal of transmitting as much task-relevant in-

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

formation as possible, given the available rate, but might be harder to make sense of otherwise.[7]

There are two conceptually distinct encoding-decoding operations that are needed in order to approach rate-distortion optima: compression, or source coding, and error correction, or channel coding (see Fig. 2). I will now review evidence that representations are very often taken to be substantially source- and channel-coded. This in turn provides evidence for the Transmission and RDC Trade-Off theses: when cognitive scientists claim that representations are typically protected from error, or efficiently trade-off richness and faithfulness, they are, implicitly, offering evidence that representations have the function of transmitting information.

4.1 Source coding

Source coding (also known as compression) is very often adduced as evidence of the presence of representations. First, researchers often explicitly appeal to compression efficiency as evidence of representational status. For example Palmer et al. (2015) argue that ganglion cells represent the future state of a variable of interest by showing precisely that these cells appear to be rate-distortion efficient, given a prediction-oriented distortion measure: ganglion cells must be representing the future, because they are keeping just that information about the past that is relevant to the future—i.e., making efficient task-relevant compression, where the task is predicting the future. As part of their case, they show that this efficient compression of the past does not simply come "from known receptive field properties" (Palmer et al., 2015, p. 6910): that is to say, a representation of the future is postulated because kinds of processing can be identified that only make sense under the assumption that the system aims at transmitting information about the future.

Palmer and colleagues explicitly rely on considerations of compression efficiency; but even among researchers that do not do so, representational systems are very often described as performing what essentially amount to source coding. Eleanor Rosch, the main figure behind the "probabilistic turn" in the research on

[7] Efficient information transmission seems to be, therefore, linked to teleofunctions—and this makes the present account a version of teleosemantics. I am happy with this identification, but recognize that it seems possible to develop the ideas in this paper in a different direction, by focusing on efficiency without taking a stand on its teleofunctional character.

One problem with this other option is that it seems to go hand in hand with taking distortion to be a free modeling parameter, one not fixed by natural selection or some other reinforcement regime. If so, any information processing system can be seen as rate-distortion efficient (or even optimal) with respect to some, perhaps ecologically implausible, distortion function. Otherwise put, it's hard to see how one can distinguish between efficient and non-efficient, optimal and merely satisficing, etc. without the help of some educated guess about the biological goals of the system. I would like to thank an anonymous reviewer for raising this point.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

concepts and categorization, famously claimed that *Cognitive Economy* is one of the two principles governing our categorization of stimuli:

> [W]hat one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible … On the one hand, it would appear to the organism's advantage to have as many properties as possible predictable from knowing any one property, a principle that would lead to formation of large numbers of categories with as fine discriminations between categories as possible. On the other hand, one purpose of categorization is to reduce the infinite differences among stimuli to behaviorally and cognitively usable proportions. (Rosch, 1999, p. 190)

A trade-off of information about the environment and the conservation of finite resources is a straightforward description of the goal of source coding. Rosch's principle of cognitive economy can be unpacked as saying that our conceptual repertoires aim at offering a comparatively small sets of signals ("to behaviorally and cognitively usable proportions") but such that decoding them downstream offers a relatively low-distortion version of the original world state ("with as fine discriminations … as possible"). The "prototype theory of conceptual representation" (Hampton, 2006, p. 79), one of the foremost accounts of concepts in cognitive psychology, directly embodies this "cognitive economy" principle: it can be shown (Martínez, 2024) that distinguishing different concepts from one another "in terms of similarity to a generic or best example" (Hampton, 2006, p. 80), which is the main tenet of the prototype theory of concepts, corresponds to optimally compressing a set of stimuli if they possess "high correlational structure" (Rosch, 1999, p. 190).

The idea that paradigmatic examples of representations are efficiently source coded is ubiquitous in cognitive science. I present without discussion a few more illustrations: in their seminal review of representations in deep learning, Bengio and colleagues claim that "good representations are expressive, meaning that a reasonably sized learned representation can capture a huge number of possibly input configurations" (Bengio et al., 2013, p. 1801). In neuroscience, the immensely influential *efficient coding hypothesis* (Barlow, 1961; Chirimuuta, 2018; see also Harpur & Prager, 1996; Sims, 2016, 2018) stems from the idea that "the current sensory situation is *represented* [by neurons] *in a concise way* which simplifies the task of the parts of the nervous system responsible for learning and conditioning." (Barlow, 1961, p. 227, emphasis added). The *predictive coding* approach to brain function (Gładziejewski, 2016; Rao & Ballard, 1999) is a very prominent development of effective-coding insights. In linguistics, work by Simon Kirby and colleagues (Kirby et al., 2015; Kirby, 2017; Smith & Kirby, 2013) aims at showing that compositional structure in languages arises "when language is under pressure to be both learnable and expressive" (Kirby et al., 2015, p. 88) where it is said explicitly that "languages which permit the formation of compressed mental representations are easier to learn than those which do not" (*ibid.*). More generally, the claim is often

made that languages look the way they do because they "must offer communicative efficiency under information processing and learning constraints" (Gibson et al., 2019, p. 389).

4.2 Channel coding

Source coding idealizes away from noise: we pretend we have a crystal clear channel, and think of how best to compress source information to squeeze it through. Channel coding takes the complementary perspective: we pretend we already have a variable, X, that optimally compresses the source, and think of how best to protect it from corruption by a noisy channel. The encoder's task is to encode our variable, and send it through the channel; the decoder's task is to decode the channel output, Y, into an estimate of the input, $\hat{X}$, correcting as much as possible for any errors that might have been introduced by the channel.

It is possible to give a general, abstract description of the way channel coding works: the trick is always to introduce redundancies in the channel input, X, so that, in effect, each value taken by this variable sits at the center of a "safety bubble" made of other values of X that the encoder will never use. This way, noise may nudge the signal away from the center of the bubble, but there is still a sufficiently high probability that it will not enter some other value's safety bubble. In general, *the objective of error correction is to identify channel inputs such that the probability of overlap between the corresponding channel outputs is suitably low.*

While source coding is an operation on the relation between signals and the source messages they are coding for, channel coding is an operation on the set of signals in and of themselves (disregarding what it is that these signals are coding for). That is to say, it should be thought of as a syntactic, as opposed to semantic, operation.

Many theorists, (e.g., Drayson, 2018; Shea, 2018), have argued that representational explanatory strategies in cognitive science necessarily rely on "vehicle realism": a commitment to entities that fall under "non-semantic types that are processed the same way by the system, and so are guaranteed to have the same content" (Shea, 2018, p. 39). A perspicuous way of thinking of vehicles (all the more so when we are dealing with emergent components of evolved systems) is as *channel-coding safety bubbles*: the way to ensure that different signals are treated reliably differently is by ensuring that if and when they are corrupted by noise they will not be nudged out of their activation-space safety bubble.

As with source coding, researchers very often rely on noise-management considerations in making judgements of representational status. For example, in *sparse coding* (Cayco-Gajic & Silver, 2019; Chalk et al., 2018; Laughlin et al., 1998; Perez-Orive et al., 2002), widely regarded as a central implementation strategy in the brain, only a comparatively low proportion of units (say, neurons) in a certain structure (say, the mushroom body, Perez-Orive et al., 2002) are active at any given time. Sparse coding is often described as a form of channel coding: for example,

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

Pérez-Orive claims that it ensures that "overlaps between individual representations are less likely than if each representation uses a large proportion of the available neurons, limiting interference" (*op. cit*, p. 364).

Synchronized neural activity (Buzsáki, 2006; Engel et al., 1992; Fries, 2005, 2015; Gray, 1999) is equally often presented, among other things, as a redundancy-based error-management mechanism. Buzsáki (2010), for example, advocates considering the assembly of synchronized cells (rather than the individual neuron) as "the fundamental unit of [neural] syntax" (*op. cit*, p. 365), among other things because "interacting assembly members … robustly tolerate noise" (*ibid.*). This noise tolerance is achieved at the cost of adding in redundancy, in the form of correlated activity of assembly cells.[8]

5 Complexity-distortion trade-offs

Information theory deals with the efficient transmission of relevant information from one variable to another, across noisy, limited-capacity channels—but it does not deal with how to *construct* these variables in the first place. At a minimum, encoding and decoding require actual, concrete systems that transform, e.g., the source variable into the channel-input variable (i.e., M into X in Fig. 1), and the channel-output variable into the destination variable (i.e., Y into $\hat{M}$). These systems might be more or less complex, and as a result capable of carrying out more of less sophisticated transformations. The third budget singled out in the RDC Trade-Off thesis has to do with the efficiency of these transformations.

Formal characterizations of computational complexity have quirks and limitations that make them less amenable to practical application in cognitive science than the formal characterizations of informational richness, or faithfulness, that we have been working with so far.[9] Still, there is a common, informal understanding of "model complexity" in cognitive science, according to which the complexity of a system can be estimated in terms of a handful of heuristics: a system is taken to be more complex, among other things, the more free parameters are needed to describe it; if it implements a nonlinear, rather than linear, transformation; or if it is recurrent, rather than feed-forward (Bassett et al., 2018; cf. Shiffrin, 2010, sec. 7).

[8] For more on the representational roles played by synchronized activity in the brain see Martínez & Artiga (2023).

[9] At least two, partially overlapping approaches are relevant here. *Classical complexity analysis* (Arora & Barak, 2009) focuses on how much time and memory is needed in order to calculate values of a given function. *Algorithmic information theory* (Li & Vitányi, 2008) studies the informational content of static objects, such as the coders in charge of variable transformations in Fig. 2. Classical complexity analysis is most useful for step-by-step algorithms over discrete sets of symbols (altough, as Rooij et al., 2019 argue, cognitive scientists still need to keep these complexity results in mind). Algorithmic information theory is a perfectly adequate framework to study the complexity of objects, but it depends on the generally non-computable notion of Kolmogorov complexity.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

Ecologically useful (e.g., sensorimotor) transformations of variables can potentially be very complex, in the above sense.[10] Complexity-distortion trade-offs are a way of managing this complexity: there will be situations in which the optimal mapping between input and output variables would require a very complex coder; but other, slightly suboptimal mappings in the vicinity are simpler—say, have fewer parameters, or less non-linearities. Choosing the simpler coders in that kind of case may make perfect engineering sense. Another piece of evidence that a system presents adaptations for information transmission is that its coders are complexity-distortion-efficient. The argument is the same as above for rate-distortion trade-offs: all variable transformations will be computationally complex to *some* degree. That is to say, the smallest model describing a system that transforms a variable A into a variable B (leaving the degenerate, $A = B$ case aside) will have *at least a few* free parameters. This is not evidence of adaptations for information processing, because it is a given for pretty much any interesting physical process, just like it is a given for pretty much any interesting physical process that it will result in nonzero mutual information between different variables. What *is* evidence for information processing (for its own sake) is that the complexity of transformations sit close to a complexity-distortion trade-off optimum—the best *prima facie* explanation of a good complexity-distortion trade-off being that the system is adapted for minimizing complexity. In the remainder of this section I discuss one important way in which tacit reliance on complexity-distortion trade-offs in the postulation of representations is manifested: the role that explicitness plays in the investigation of representations.[11]

A major link between computational complexity and representational status is the idea, popular in computational neuroscience, that whether some quantity or property counts as *explicitly represented* by a population of neurons has to do with whether it can be *easily decoded* from that population.[12] For example, according to Kriegeskorte and Diedrichsen, an explicit representation is "a representation … in a format that enables it to be decoded *in a single step* by biological neurons" (2019, p. 411, my emphasis). Similarly, in Hong et al. (2016), explicit information is identified with "easily accessible information … as measured via the performance of linear classifiers …" (613, see also DiCarlo et al., 2012, p. 417).

[10] But perhaps need not be, as ecological and embodied cognitive scientists remind us. See Chemero (2011); Wilson & Golonka (2013), and §6 below.

[11] There are other ways in which complexity management is implicitly or explicitly exploited in cognitive science. Compositionality is a particularly important one, but a full discussion of this topic is matter for another paper.

[12] What neuroscientists mean by "explicit representation" is very close to what philosophers of cognitive science mean by "representation" *simpliciter*. As Poldrack (2021, p. 2) points out, the operationalized notion of "representation" (*simpliciter*) in neuroscience corresponds to all and any "systematic relationship between the activity of neurons and the structural … features of the world". This is extremely liberal, and in particular means that any neural variable A represents any worldly variable B if $I(A;B) \neq 0$.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

In fact, ease of decoding is already implicit in Marr's famous dictum that "[a] representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this." (Marr, 2010, p. 20): he points out that the Arabic representation of numerals makes "the number's decomposition into powers of 10" explicit, while their binary representation makes "the number's decomposition into powers of 2" explicit (*ibid.*). This can be straightforwardly interpreted as the observation that a decoder that extracts a decomposition of, e.g., 37 into powers of 10 from its decimal representation 37 is algorithmically trivial, and an algorithm that extracts its decomposition into powers of 2 from its binary representation 100101 is equally trivial; but that doing it the other way around is less trivial.

The line of thought that goes from representational status to the explicit presentation of information, to presentations of information that are easy to decode, I suggest, relies (sometimes tacitly, sometimes rather explicitly) on something very much like the Complexity leg of the RDC Trade-Off thesis: ease of decoding provides evidence of an adaptation for information processing because the most straightforward explanation of an easily decodable signal is that it was purposely formatted so as to be handled by decoders of that sort. The emphasis on explicitness (understood as ease of decoding) in the attribution of representational status is a tacit recognition of this fact.

To see how considerations of computational complexity play an important role in the attribution of representational status to neural activity, think of the primate ventral visual processing stream (the processing path that goes from V1, to V2, to V4, to the inferotemporal cortex [IT], and which underlies object recognition among other things—see DiCarlo et al., 2012, fig. 2), operating in the context of an object-recognition task. That is to say, the subject is confronted with a visual scene, and asked to decide whether, e.g., a car is present or not in it. One question we may want to ask is, where in the ventral processing stream, from V1 to IT, is this decision represented? One way to answer this question, which theorists educated in the kind of information-as-correlation views discussed in §3 might find appealing, is the following:

1. Define the *ideal destination variable* as a binary random variable which takes value 1 (0) if there is a (no) car in the visual scene under consideration.

2. Find the neuronal population, P, whose activity has the highest mutual information with the ideal destination variable.

3. The car/no car decision happens in whatever processing stage P is.

That is to say, according to this broadly "correlational" heuristics, the decision happens there where neurons carry the most information about the decision. It is instructive to see how this heuristics is bound to give the wrong answer—or at least only coincidentally give the right one: *as we ascend from the retina to IT, information about the ideal destination variable can only be lost.* This fact follows

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

from an information-theoretic result called the *data processing inequality* (Cover & Thomas, 2006, p. 35): given how the object-recognition task works, *all* of the information relevant to making the decision in question is, of necessity, in the pixels of the visual scene that the subject is exposed to. Processing and transforming this information cannot magically make it increase. This means that the earlier the processing stage is, the more information it will carry about the presence or absence of cars—and about any other thing the source pixels might carry information about.

If it is not about increasing information about the decision, and it cannot be, then what is ventral processing all about? The answer that can be gleaned from the work of DiCarlo and colleagues is that it is about *constructing a signal that can be read by a low-complexity decoder*: the ventral stream is seen as trading off distortion (which increases as we move up, thanks to the data processing inequality) with complexity (which decreases), up to the point at which the decision can be linearly decoded. Linear decodability, researchers assume, is simple enough, and the decision counts as explicitly represented.

Many other researchers rely on similar complexity-dependent ideas in their analyses of the representational features of the ventral stream. For example, Brouwer and Heeger (2009, 2013) report how color can be decoded better from V1 than from V4, but that only the color representation in V4 "reflects perceptual color space" (2009, p. 13992). By this they mean that, in V4, "perceptually similar colors [evoke] the most similar responses" (*ibid.*)—that is to say, again, there is a low-complexity, monotonic translation between evoked response and perceptual color.[13]

In my formulation of the Transmission thesis (§1) I claimed that representations are primarily vehicles of information transmission. There, "primarily" was signaling the fact that cognitive scientists often also recognize other, not signal-like, kinds of representations: e.g., cognitive maps (Epstein et al., 2017), memories (Robins, 2016), or generative models (Friston, 2010). We can think of these mostly static representations as being part of, or identical with, the coders in charge of constructing and transforming the dynamic representations, the signals, that are at the core of the adaptive generation of behavior. This is the sense in which the latter are primary, and the former subordinated to them. See Martínez (forthcoming) for more on these so-called *structural representations*.

[13] Ritchie et al. (2020) have persuasively argued that the fact that one can linearly decode a variable of interest, V, from fMRI data is not good evidence that V is represented in the neural population to which those data correspond.

The claim I am substantiating in this section (as part of my defense of the RDC Trade-Off thesis) is that paradigmatic representations trade off distortion and complexity. I am offering the fact that cognitive scientists identify explicitness in representation with ease of decoding as evidence. This is of course compatible with the inference from ease of fMRI decoding to representational status being problematic in various ways: fMRI data patterns are a fallible guide to neural activation patterns (*ibid.*); or it may just so happen that some neural activation patterns can be linearly decoded, by chance or for unrelated reasons, but do not constitute signals because no brain decoder actually consumes them (see Rathkopf, 2017 for related points.)

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

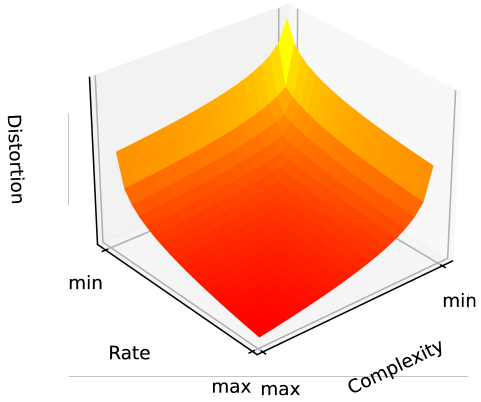

Figure 5: An idealized representational surface

6 The representational surface

In the fully general case both rate-distortion and complexity-distortion trade-offs are important. First, getting information from one variable to another involves a decision about how rich the signals exchanged should be; e.g., do we need to send a full retinotopic map downstream, or will a sketchier representation of the gist of the visual scene do, for the purposes at hand? This is the rate-distortion trade-off. But also, second, how complex can our coders afford to be? Very sophisticated coders (e.g., ones that have access to, or themselves qualify as, a rich model of the environment; or ones able to perform highly non-linear transformations on the incoming signal) can squeeze out the most information from signals, at a high memory and processing cost. Simpler coders can squeeze from the incoming signals *some* information relevant to the task at hand—and that might be enough. This is the complexity-distortion trade-off.

I suggest that the Pareto frontier (see §1) of rate, distortion, and complexity minimization, a two-dimensional surface in the space spawned by these three quantities, is an important object to keep in mind when reasoning about representations in cognitive science. Fig. 5 offers an idealized example: a surface where distortion monotonically decreases as rate and complexity increase. When complexity or rate are very low, distortion is near its maximum. As both increase to their maximum, distortion approaches its minimum.

Many recalcitrant debates about the role that representations play in cognitive function can be best understood as disagreements about the shape of the representational surface, and about the position that representations occupy at or near it. For example, a very popular idea in enactive, ecological, embodied approaches to cognitive science is that the world is rich enough in information on its own, which means that there is no need for "cognitive enrichment" (Wilson & Golonka,

2013, p. 2) to be performed over internal models and representations: "the world is its own model" (McClamrock, 1995, p. 114); "the natural environment is already rich with affordances and information that can guide behavior" (Chemero, 2011, p. 27). But, of course, the very idea of cognitive enrichment of incoming information is misguided: all of the information necessary for performance of a task *has* to come from an (extra-cognitive) source. Cognitive systems cannot magically enrich the source, in the sense of putting in information that was not there (§5). The environment is, quite literally, as "rich with affordances and information" as it gets.[14] The dispute is more accurately presented as being about the computational complexity of extracting this information in a way that makes it usable for the production of behavior. Classical cognitive scientists labor under the assumption that the computational complexity of this process is high; 4E cognitive scientists argue that it is low. This is explicit in Chemero (2011, Chapter 6), where he argues that certain important perceptual tasks can be undertaken by two-layer artificial neural networks (a low complexity computational device), and therefore "without manipulating representations" Chemero (2011, Chapter 6).

The idealized surface in Fig. 5 is symmetrical on the contribution that rate and complexity make to reducing distortion. 4E cognitive scientists can be read as making the bet that the representational surface for ecologically relevant tasks looks more similar to Fig. 6: complexity "saturates" very quickly, and after that distortion is dominated by rate. That is to say, the source (the world) already comes motor-coded, or nearly so, and adaptations for information transmission are unnecessary.

Whether the representational surface looks like Fig. 5 or like Fig. 6 for any given task is a thoroughly empirical question: a question about achievable trade-offs in multiobjective optimization.

7 Concluding remarks

I have argued that carrying information from source to destination is a substantial engineering achievement: protecting a signal from noise, choosing an encoding that makes the most of the available channel (given the uses the signal will be put to), all the while keeping the complexity of coders as low as possible, cannot be realistically done without goal-directed adaptations—without biological functions. Moreover, I have also argued that these adaptations map onto the properties that we associate with representations. I have reviewed some examples here:

[14] While I see the stance on "cognitive enrichment" in the main text as the most natural way to apply the lessons of information theory to this question, I recognize that the matter is to a large extent terminological: one could very well say that predicting, or reconstructing, the value of an extramental variable from other cues counts as information enriching. My preferred way of describing this situation is as one in which the latter cues carry (enough) information about the former extramental variable for whatever purposes it is to be put downstream. This is, I believe, how things are set up in Shannon's theory of information, and it is also close to what Gibson means by "ecological information"; but talk of enrichment may make perfect sense in other contexts and for other purposes.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

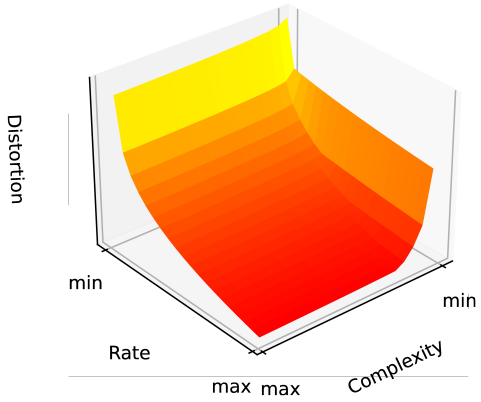

Figure 6: A 4E-friendly representational surface

noise protection is related to the presence of representational vehicles; compression is related to the widespread idea that representations aim at being "cognitively economical" in Eleanor Rosch's sense; computational complexity management is behind the notion of explicit representation in neuroscience. Other connections between efficient RDC trade-offs and representational status exist; mapping and discussing them is matter for another time.

Because entities with the kinds of properties we associate with representations emerge as solutions to multiobjective (rate-distortion-complexity) optimization problems, "representational surfaces" are useful mental models when thinking of representational phenomena. Sketching, even if qualitatively, the shape of the problem to be solved (that is to say, the Pareto frontier in RDC space) can fruitfully inform investigation of the role that representations play in cognition. As I have briefly discussed in §6, whether representations will emerge as efficient solutions to some RDC trade-off problem will depend on the shape of the representational surface—a thoroughly empirical matter, but one in which a number of formal approaches to inference and cognition can help. For example, the *Minimum Description Length Principle* (commonly abbreviated as MDL, Barron et al., 1998; Grünwald, 2007) identifies the best model of some data, roughly, as the one that allows to express these data as succintly as possible. One can think of the "optimal universal codes" (Grünwald, 2007, Chapter 6) that do this as, first, encoding a model and, second, encoding individual data given that model. These two stages are closely related to the codecs and signals, respectively, discussed in this article (see Grünwald & Vitányi, 2003). MDL codes sit on the representational surface, and, as such, they can help guide our discovery of representations in computational cognitive science. Two things should be noted, though. First, as Barron et al. (1998, p. 2744) remark, "there is no 'mechanical,' i.e., algorithmic, way to find

the 'best' model of data among all computable models": heuristics-based, informal approaches to RDC tradeoffs of the sort discussed in this paper are unlikely to be superseded by MDL. Second, MDL aims at lossless reproduction of data (although see Grünwald, 2007, sec. 17.8), while, as I have argued, most interesting exercises of information transmission in cognitive systems are lossy—i.e., do not aim at zero distortion.

Another prominent, potentially useful framework in this connection is the *Free Energy Principle* (commonly abbreviated as FEP, Friston, 2012; Parr et al., 2022), which promotes a variational-Bayes-based approach to inference in partially observable Markov processes, and other similar ones, perhaps in continuous as opposed to discrete time. Many of the ideas in the FEP program are compatible with, but go beyond, the focus on information- and complexity-based properties of information-processing systems that I have advocated for here: one can get behind the idea that cognition aims at managing RDC costs in a satisficing manner without necessarily endorsing the very ambitious idea that life and mind are essentially processes of variational Bayes optimization. For an approach to representation that draws on rate-distortion theory and touches on variational inference, see (De Llanza Varona et al., 2024; see also De Llanza Varona & Martínez, 2024).

Another use for representational surfaces, that should be explored elsewhere, is charting where on the surface we should expect to find different kinds of representational entities. E.g., plausibly, concepts are lower rate, lower complexity entities than perceptual representations; and the ventral visual-processing stream could be seen as tracing a tracing a path along the representational surface, where distortions increase, but rate and complexity are greatly reduced.

Acknowledgments
I would like to thank the Buenas Migas work in progress group, audiences and commentators at many talks, my three reviewers for this journal, and many more for a few other journals.
This work has been funded by the Spanish Ministry of Science and Innovation, through grants PID2021-127046NA-I00 and CEX2021-001169-M (MCIN/AEI/10.13039/501100011033); and by the Generalitat de Catalunya, through grant 2021-SGR-00276.

References

Akagi, M. (2021). Cognition as the sensitive management of an agent's behavior. *Philosophical Psychology*, *0*(0), 1–24. https://doi.org/10.1080/09515089.2021.2014802

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Publishers.

Arora, S., & Barak, B. (2009). *Computational complexity*. Cambridge University Press.

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6, 6), 359–371. https://doi.org/10.1038/s41583-021-00448-6

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *The British Journal for the Philosophy of Science*, *4*(14), 147–157. https://www.jstor.org/stable/685989

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*, 217–234.

Barron, A., Rissanen, J., & Bin Yu. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*(6), 2743–2760. https://doi.org/10.1109/18.720554

Bassett, D. S., Zurn, P., & Gold, J. I. (2018). On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience*, *19*(9, 9), 566–578. https://doi.org/10.1038/s41583-018-0038-8

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Berger, T. (2003). Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience*, *2*(11), 947. https://doi.org/10.1038/14731

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X19000049

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-m

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Brouwer, G. J., & Heeger, D. J. (2013). Categorical clustering of the neural representation of color. *The Journal of Neuroscience*, *33*(39), 15454–15465. https://doi.org/10.1523/JNEUROSCI.2472-13.2013

Buzsáki, G. (2006). *Rhythms of the brain*. Oxford University Press.

Buzsáki, G. (2010). Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron*, *68*(3), 362–385. https://doi.org/10.1016/j.neuron.2010.09.023

Cayco-Gajic, N. A., & Silver, R. A. (2019). Re-evaluating circuit mechanisms underlying pattern separation. *Neuron*, *101*(4), 584–602. https://doi.org/10.1016/j.neuron.2019.01.044

Chalk, M., Marre, O., & Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, *115*(1), 186–191. https://doi.org/10.1073/pnas.1711114115

Chemero, A. (2011). *Radical embodied cognitive science*. MIT press.

Chirimuuta, M. (2018). The development and application of efficient coding explanation in neuroscience. In A. Reutlinger & J. Saatsi (Eds.), *Explanation beyond causation: Philosophical perspectives on non-causal explanations* (pp. 164–184). Oxford University Press. https://doi.org/10.1093/oso/9780198777946.003.0009

Chollet, F. (2021). *Deep learning with python*.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. New York: Wiley.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Computational Neuroscience Series.

De Llanza Varona, M., Buckley, C., & Millidge, B. (2024). Exploring action-centric representations through the lens of rate-distortion theory. In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, T. Verbelen, & M. Wisse (Eds.), *Active inference* (Vol. 1915, pp. 189–203). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47958-8_12

De Llanza Varona, M., & Martínez, M. (2024). Synergy makes direct perception inefficient. *Entropy*, *26*(8), 1–22. https://doi.org/10.3390/e26080708

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Drayson, Z. (2018). The realizers and vehicles of mental representation. *Studies in History and Philosophy of Science Part A*, *68*, 80–87. https://doi.org/10.1016/j.shpsa.2018.01.005

Dretske, F. (1981). *Knowledge and the flow of information*. The MIT Press.

Dretske, F. (1988). *Explaining behavior. Reasons in a world of causes*. The MIT Press.

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences*, *15*(6), 218–226. https://doi.org/10.1016/0166-2236(92)90039-B

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, *20*(11, 11), 1504–1513. https://doi.org/10.1038/nn.4656

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474–480. https://doi.org/10.1016/j.tics.2005.08.011

Fries, P. (2015). Rhythms for cognition: Communication through coherence. *Neuron*, *88*(1), 220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127. https://doi.org/https://doi-org.sire.ub.edu/10.1038/nrn2787

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy*, *14*(11, 11), 2100–2121. https://doi.org/10.3390/e14112100

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. https://doi.org/10.1007/s11229-015-0762-9

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, *24*(1), 31–47. https://doi.org/10.1016/S0896-6273(00)80820-X

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D., & Vitányi, P. M. B. (2003). Kolmogorov complexity and information theory. With an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, *12*(4, 4), 497–529. https://doi.org/10.1023/A:1025011119492

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of Learning and Motivation*, *46*, 79–113. https://doi.org/10.1016/S0079-7421(06)46003-5

Harpur, G. F., & Prager, R. W. (1996). Development of low entropy coding in a recurrent network. *Network (Bristol, England)*, *7*(2), 277–284. https://doi.org/10.1088/0954-898X/7/2/007
Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247
Isaac, A. M. C. (2019). The semantics latent in Shannon information. *The British Journal for the Philosophy of Science*, *70*(1), 103–125. https://doi.org/10.1093/bjps/axx029
Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, *24*(1), 118–137. https://doi.org/10.3758/s13423-016-1166-7
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016
Körding, K. P., & Wolpert, D. M. (2004). The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences*, *101*(26), 9839–9842. https://doi.org/10.1073/pnas.0308394101
Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual Review of Neuroscience*, *42*, 407–432. https://doi.org/10.1146/annurev-neuro-080317-061906
Laughlin, S. B., de Ruyter van Steveninck, R. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, *1*(1), 36–41. https://doi.org/10.1038/236
Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications. Texts in computer science* (Vol. 9). Springer, New York,.
MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
Mann, S. F. (2023). The relevance of communication theory for theories of representation. *Philosophy and the Mind Sciences*, *4*. https://doi.org/10.33735/phimisci.2023.10992
Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press.
Martínez, M. (2019a). Deception as cooperation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *77*, 101184. https://doi.org/10.1016/j.shpsc.2019.101184
Martínez, M. (2019b). Representations are rate-distortion sweet spots. *Philosophy of Science*, *86*(5), 1214–1226. https://doi.org/10.1086/705493
Martínez, M. (2024). The information-processing perspective on categorization. *Cognitive Science: A Multidisciplinary Journal*, *48*(2), e13411. https://doi.org/10.1111/cogs.13411
Martínez, M. (forthcoming). Structural representation as complexity management. In G. Piccinini (Ed.), *Neurocognitive foundations of mind*.
Martínez, M., & Artiga, M. (2023). Neural oscillations as representations. *The British Journal for the Philosophy of Science*, *74*(3), 619–648. https://doi.org/10.1086/714914
McClamrock, R. A. (1995). *Existential cognition: Computational minds in the world*. University of Chicago Press.
Millikan, R. G. (1993). What is behavior? A philosophical essay on ethology and individualism in psychology, part 1. In *White queen psychology and other essays for alice* (pp. 135–150). The MIT Press. Bradford Books.
Myerson, R. B. (1997). *Game theory: Analysis of conflict*. Harvard University Press.
Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT Press.
Noë, A. (2004). *Action in perception*. MIT Press. https://books.google.com?id=kFKvU2hPhxEC
O'Connor, C. (2020). *Games in the philosophy of biology*. Cambridge University Press. https://doi.org/10.1017/9781108616737
Palmer, S. E., Marre, O., Berry, M. J., & Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, *112*(22), 6908–6913. https://doi.org/10.1073/pnas.1506855112
Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. The MIT Press.
Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., & Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science*, *297*(5580), 359–365. https://doi.org/10.1126/science.1070502
Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1–2), 1307–1325. https://doi.org/10.1007/s11229-020-02793-y
Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, *10*(3, 3), 173–185. https://doi.org/10.1038/nrn2578
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79. https://doi.org/10.1038/4580
Rathkopf, C. (2017). Neural information and the problem of objectivity. *Biology & Philosophy*, *32*(3), 321–336. https://doi.org/10.1007/s10539-017-9561-7
Rieke, F., Warland, D., Van Steveninck, R. de R., & Bialek, W. S. (1999). *Spikes: Exploring the neural code*. The MIT Press.
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2020). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axx023
Robins, S. (2016). Representing the past: Memory traces and the causal theory of memory. *Philosophical Studies*, *173*(11), 2993–3013. https://doi.org/10.1007/s11098-016-0647-x
Rooij, I. van, Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.
Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 189–206). The MIT Press.
Sayood, K. (2017). *Introduction to data compression*. Morgan Kaufmann. https://books.google.com?id=3DFHDgAAQBAJ
Scarantino, A. (2015). Information as a probabilistic difference maker. *Australasian Journal of Philosophy*, *93*(3), 419–443.

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.
Shannon, C. E., & Weaver, W. (1998–1949). *The mathematical theory of communication*. University of Illinois press.
Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.
Shepherd, J. (2021). *The shape of agency: Control, action, skill, knowledge* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198866411.001.0001
Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, *2*(4), 736–750. https://doi.org/10.1111/j.1756-8765.2010.01092.x
Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.
Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656. https://doi.org/10.1126/science.aaq1118
Skyrms, B. (2010). *Signals: Evolution, learning & information*. New York: Oxford University Press.
Smeets, J. B., & Brenner, E. (1999). A new view on grasping. *Motor Control*, *3*(3), 237–271.
Smith, K., & Kirby, M. T. S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35), 1348–1353.
Sprevak, M. (2019). Two kinds of information processing in cognition. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-019-00438-9
Timme, N. M., & Lapish, C. (2018). A tutorial for information theory in neuroscience. *eNeuro*, *5*(3), ENEURO.0052–18.2018. https://doi.org/10.1523/ENEURO.0052-18.2018
Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9, 9), 907–915. https://doi.org/10.1038/nn1309
Varela, F. J., Thompson, E., & Rosch, E. (2016). *The embodied mind: Cognitive science and human experience* (revised edition edition). The MIT Press.
Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, *37*(4), 681–691. https://doi.org/10.1016/S0896-6273(03)00035-7
Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105
Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00058

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321