CellPress

Trends in
Cognitive Sciences

Feature Review

Three aspects of representation in neuroscience

Ben Baker, [1],* Benjamin Lansdell, [2] and Konrad P. Kording [3]

Neuroscientists often describe neural activity as a representation of something, or claim to have found evidence for a neural representation, but there is considerable ambiguity about what such claims entail. Here we develop a thorough account of what 'representation' does and should do for neuroscientists in terms of three key aspects of representation. (i) Correlation: a neural representation correlates to its represented content; (ii) causal role: the representation has a characteristic effect on behavior; and (iii) teleology: a goal or purpose served by the behavior and thus the representation. We draw broadly on literature in both neuroscience and philosophy to show how these three aspects are rooted in common approaches to understanding the brain and mind. We first describe different contexts that 'representation' has been closely linked to in neuroscience, then discuss each of the three aspects in detail.

Highlights

Recent scientific discourse has pointed to challenges in the way the terms 'representation' and 'code' are commonly used in neuroscience.

In neuroscience, one current line of thinking focuses on the kinds of representations formed in neural systems trained to predict incoming sensory inputs.

Machine learning researchers are interested in how (artificial intelligence) AI can acquire representations of causal variables and representations that are disentangled.

Philosophers work on questions about how representations relate to psychological explanation, to knowledge, and to phenomenological experiences.

The best way to study neural representations and their impact on the mind and behavior is a multifaceted and interdisciplinary topic and we work here to build bridges between major sub-areas interested in this question.

Why and how to reassess representation in neuroscience

Experimental data in neuroscience are interpreted and communicated in terms of certain key concepts, including computation, code, and representation. Patterns of neural activity are often referred to as representations, however, there is no widely agreed upon definition of 'representation' and the implied meaning of the term across different subfields and individual scientists is somewhat vague and inconsistent. Recent theoretical work has pointed out a disconnect between the kind of understanding suggested by claims involving neural representation and the kind of observations typically offered as evidence of representations [1–5]. A survey of authors within two areas of neuroscience research revealed significant variation in respondents' ways of describing what is implied by 'representation' [6]. Neuroscience is still a rapidly growing field and has substantial intellectual exchange with philosophy, psychology, cognitive science, and artificial intelligence (AI), where others have pointed out equivocal usage of 'representation' and related concepts [7,8]. Given the lack of theoretical unity around a conception of representation in neuroscience, coupled with the increasing technological efforts to probe and understand the brain, it is essential that the field develop a clearer understanding of what the notion of 'representation' does and should do for neuroscientists. This motivates our paper, which steps back to assess a cluster of common ideas around representation in neuroscience and carefully formulates an account for reference in future work.

In reappraising the notion of representation, we draw considerably on work by philosophers. Representations are a focal topic in contemporary philosophy of mind and, in some form, going back to the ancient period. While philosophers have offered many accounts of representation that differ in their details, a few broad themes can be distilled from the theoretical challenges that groups of philosophers have posed and tried to answer. These philosophical exchanges concern the assumptions that are obscured in how the term is currently used in neuroscience, but for various reasons usage in neuroscience has been largely detached from that in philosophy

[1]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Neuroscience, Bioengineering, University of Pennsylvania, CIFAR, Philadelphia, PA, USA

*Correspondence:
tbak@sas.upenn.edu (B. Baker).

https://doi.org/10.1016/j.tics.2022.08.014

[9]. It is one aim of this paper to alleviate the relative lack of cross-disciplinary discourse about representation.

With a murky collective conception of representation, neuroscientists may miss the import of each others' claims and fail to coordinate their views about what kind of research is revelatory or worthy of pursuit. As things stand, two neuroscientists may nominally agree that some neural activity represents some feature *F*, but deeply disagree about what (if anything) this tells us about larger perceptual or cognitive processes. Or they may nominally disagree about whether some neural activity represents *F* rather than *G*, while looking at the same findings. Resolving such disagreements is not simply a matter of conducting another experiment, but rather of considering as broad a range of neuroscientific experiments as possible, including hypothetical ones, alongside basic premises about cognition and behavior and about what counts as good explanations of them. This points to why, in order for the scientific community to more effectively pursue the discoveries about the brain that it sees as most valuable, it can help to attend to the way philosophers have charted the conceptual landscape around representation. We work here to bridge this gap in part by drawing broadly on philosophical literature and applying it to paradigmatic examples from current neuroscience. By these means we aim to disambiguate the notion of representation in a way that can advance the communication and research goals of neuroscientists.

Our account distinguishes three aspects of representation in neuroscience: three conditions entailed by the identification of representations when they are used in a characteristic kind of explanation of how (some part of) the brain works. Briefly, the first aspect is a relation of correlation between a representation and the thing it represents, the second aspect relates a representation to behavior via a causal role, and the third aspect puts representation in a teleological (purposive) context. We highlight how each of these three is brought out by common approaches in philosophy and how each is implicated by the explanatory significance that representations are attributed in common neuroscientific usage. We proceed as follows: in the first section, we discuss the range of meanings that 'representation' has been most closely linked to in neuroscience, linking it to the sorts of evidence it is based on and the sorts of claims it is used to support; in the second section, we discuss each aspect in turn, putting each in the context of philosophical views that define or emphasize it, and illustrating each aspect in light of experimental examples.

Common uses of 'representation' in neuroscience

On a first pass, a neuroscientist might define a representation as a state within the brain that stands in for some relevant feature of the world, acting as a token for that feature [10]. This definition suggests both that the representing states within the brain relate to the things being represented and that these states are used by the animal; they interact with other brain processes in a way that relates to the animal's behavior. Thus, identifying a neural representation of some object *F* is supposed to help us understand how the animal whose brain it is does interesting, *F*-related things.

There are at least three distinct contexts in which neuroscientists commonly use the concept of representation. One is in relating neural activity to inputs by measuring a statistical correspondence between neurons (or populations thereof) and salient features of an animal's environment. This perspective starts from a statistical relationship between neural activity and the environment and infers the relationship between a representation and a thing represented. Neuroscience experiments often identify patterns of covariation between neural responses and some experimental variable, either stimulus parameters or quantified behavior, the analysis of which usually involves tuning curves [10] (Box 1). It is easiest to quantify correlations to things in the world

Box 1. Tuning curves, codes, and internal variables

Many neuroscientists further interpret correlational experiments in terms of a communication metaphor, which can be quantified with Shannon's information theory [83]. This approach interprets neural activity as transmitting one among multiple possible signals that are sent through presynaptic connections and received by postsynaptic neural structures. This interpretation allows for asking *how much* information about a given external variable is contained in the activity [84]. The neurally transmitted information is then interpreted as an 'encoded' version of features of the outside world. This notion of a 'neural code' assumes that the relevant information from the environment is transduced by peripheral sensory mechanisms and encoded into the format in which the neurons communicate and assumes further that this neural activity is subsequently decoded by downstream processes in the brain. The notion of a neural code is ubiquitous in neuroscience and may often be used interchangeably with that of neural representation [85–87].

In the context of thinking about algorithms underlying computations achieved by the brain, the contents of some representations may not pertain to external environmental variables at all. Rather, neural activity might represent something about the state of the system that carries out the algorithm. As an example, consider the ramping of spiking activity in the lateral intraparietal cortex (LIP) when primates are presented with multiple-moving-dot stimuli, the average motion of which is ambiguous between directions. According to one theory, the spiking rate of neurons in LIP represents the integrated evidence for the stimuli moving in one direction or the other [88–90]). This theory may be off the mark, but it exemplifies the way neurons may be taken to represent something about the computational system itself, like how much evidence it has accrued, rather than representing something about an external object.

that an animal is currently experiencing, so this approach to neural representation is most often associated with work on early sensory areas, where activity largely only depends on incoming stimuli, providing the cleanest connection between neural activity and features in the world. Neuroscientists studying vision, for instance, may record from neurons in V1 and observe how their activity is correlated with visual stimuli it was concurrently presented with. In short, the neural activity may simply be taken to be a representation of whatever environmental variable the activity is found to correlate well with.

A second common context for thinking about representations is in relating neural activity to behaviors based on observations of certain kinds of animal behavior. The aforementioned informal definition suggests that an animal uses representations in the production of adaptive behavior, supporting the idea that an animal's behavior can tell us something about things represented in parts of the animal's brain. Specifically, since representations are presumed to serve flexible and temporally extended behaviors, behavior that is not an immediate response to stimulation seems indicative of representation [5,11]. For example, if an animal possesses a representation of a predator that is not in range of sensory systems, it can use this representation to make predictions about the predator's future location and act accordingly. Note that the natural focus in this context is *not* on early sensory areas with activity tightly tied to incoming stimuli. Animals whose behavior is largely reactionary do not invite this form of inference about internal representations driving the animal's clever response to distant parts of the world. As some philosophers would put it, some behavioral tasks are not 'representation-hungry', meaning they do not call for explanation in representational terms [12]. Although observations of 'representation-hungry' behavior do not, by themselves, speak to how an animal's nervous system implements the relevant representations, it is in the context of such observations that many neuroscientific hypotheses are formed about what an animal's brain might represent [13].

A third context in which neuroscientists often use the idea of representation is in talking about different kinds of algorithms that might support general capabilities that systems neuroscientists are interested in, like decision making, memory, and abstract reasoning. Marr offered a way of thinking about such algorithms in relation to the larger (especially visual) computations that depend on them and to their physical implementation in (especially neural) hardware [14]. (Some philosophers and scientists have used a different notion of computation, which is defined just as a rule-governed manipulation of symbols, irrespective of any semantic content they may have [15–19].) The contents of the representations in such an algorithm may appear to have nothing directly to do with the environment or with behavior, but mainly with the internal states of the system itself as its algorithm is playing out over time. For example, an important variable to represent, in some algorithms, could be how much evidence has been accrued for some possibility, or how much remains of some internal resource. More generally, neuroscientists thinking about how a brain (or some of its parts) might solve a certain problem often rely importantly on a notion of representation that is neither associated with sensory activity *per se*, nor with any particular behavioral response to part of the world beyond immediate sensory access.

We have just illustrated three contexts in which neuroscience uses the concept of representation: one based on behavior, one based on relations between neural activity and the world, and one used in the exploration of computational algorithms in the brain. These approaches suggest different forms of evidence for representations and each has different motivating assumptions. A behavior-based approach may make no commitments about how a given representation is implemented. An approach based on neural correlates may come with unstated assumptions about the role of such correlates in shaping behavior, but may not directly examine that role. A computation-based approach revolves around the issue of what algorithms may be implemented in a nervous system. Despite these differences, often neuroscientists may mean to refer to the same underlying phenomenon across these contexts. The result of these heterogeneous and unarticulated assumptions is that an imprecise notion of representation pervades neuroscientific communications.

Later we carefully articulate a view of representations in terms of three aspects. One may note thematic links between each of the three aforementioned contexts and each of the three aspects we describe, but our point is decidedly not that each aspect only concerns certain investigative approaches. We will describe each aspect in general terms that are relevant to all of the earlier neuroscientific contexts. Our goal is both to illuminate an intended meaning that typical usage in neuroscience has been circling around and to enable clearer claims about representation in neuroscience going forward.

Drawing on philosophical perspectives

Representation is a core issue in much contemporary work in philosophy of mind and some of the thinking in neuroscience is plausibly already influenced by philosophical conceptions of representation. (The movement of central ideas across disciplines is a difficult question we will not directly address here.) The scope and diversity of philosophical literature on representation is far too great for us to succinctly summarize here, but we have drawn widely from works that make the clearest connections to the physiology or functional description of animals or artificial systems. We also point out roots or analogs of these perspectives in much older philosophical work and apply them in the context of concrete neuroscientific examples.

In the broadest terms, philosophers commonly understand representations to be entities that have semantic content (i.e., a representation is *about* something), where different representations can represent the same thing in different ways. For example, a map of New York, a painting of New York, and an utterance of ‘New York’ could all represent the same place (content) in different ways. Representational views of the mind, and the closely related notion of intentionality, have a long history of discussion in philosophy [20,21]. (This kind of commitment to the reality of representations as tokened in internal states can be contrasted with a view where the ‘intentional stance’ we take toward certain systems is not taken to depend on the workings of underlying physical components [22].) In the era of brain research, there have been various philosophical accounts of what representation involves, which generate different assessments of the way neuroscience experiments try to reveal representations [23–34].

We build on these accounts and critiques to distinguish three aspects of representation that are implicated in the explanatory role they are commonly asked to play in neuroscience. (i) Correlation: a neural representation correlates to its represented content; (ii) causal role: the representation influences behavior; and (iii) teleology: a goal or purpose should be furthered by the behavior and thus the representation. We illustrate that the prevailing neuroscientific use of the term 'representation' seems concerned with all three aspects to some extent, although it is often ambiguous with respect to the latter aspects. We work toward more clarity around this issue by examining each of these aspects of representation and elaborating their significance in connection to both philosophy and neuroscience.

Note that we are *not* proposing that neuroscientists should all be focused on identifying representations as conceived here or be equally interested in investigating all three aspects. Different parts of the field are in positions to further our knowledge about different aspects of representation to different degrees, and some of what we want to understand about the brain is not a matter of what some part of it represents. Our intent is rather to carefully spell out what representations entail, given the explanatory role they have in common neuroscientific usage.

Aspect 1: correlation

The first aspect describes representations as matching or fitting to the things they represent, in the sense of correlating to them. For instance, it could be partly by having inner states that correlate to red fruit in an animal's environment that the animal can recognize such fruit. In general, this aspect says that *if some neural activity N represents some event or feature of the world F, then one should be able to find evidence of a correlation between N and F.*

The significance of correlation is deeply rooted in philosophical views of representation (Box 2). The bulk of the philosophical literature on the topic suggests that correlation does not suffice for representation, as some further theoretical principles are needed to determine whether neural activity is a representation and of what [24,31,35–38]. The first aspect is insufficient for representation because, in short, correlations are sure to exist between neural processes and all sorts of things they are not representations of. For instance, circadian rhythms can be found in neurons

Box 2. Classic correlations

A classic example of representations hypothesized on the basis of correlational evidence is Hubel and Wiesel's [91] original finding of receptive fields associated with cells in the visual cortex of a cat. In brief, these cells' activity were shown to correlate to stimuli in specific portions of the cat's visual field via experiments where cell responses and stimulus properties were simultaneously recorded. This line of work revealed how features of the animal's visual environment are reflected in the organized activity of certain parts of the brain, making a huge contribution to our understanding of the neural processes that underlie visual perception and behavior more broadly.

Correlations are also prominent in a way of thinking about representations deeply rooted in western philosophy. Often what philosophers discuss is not merely correlation but a structural correspondence or isomorphism, where internal relationships among parts of the represented entity *F* should match internal relationships among parts of the internal activity *N* (we describe Aspect 1 in terms of correlation for simplicity, but one could substitute in a form of structural correspondence with little change to the overall account). A prevailing view among European philosophers in the early modern period was a resemblance-based view of representation. For example, Locke and Hume conceived of internal 'impressions' or 'ideas' of external objects as representing those objects in virtue of resembling them [92,93]. Considerably farther back in the history of philosophy, in *De Anima*, Aristotle conceives of registration by perceptual organs on analogy to the symbol that remains in the shape of some wax after a signet ring has been pressed into it (Book II, Chapter 12) [94]. Philosophers have found compelling the notion that intelligent behavior depends on internal states whose shapes and interactions mirror things in the outside world long before any were aware of the brain's capacities to implement complex formal relationships.

and other cells throughout the brain, but this does not make these cells representations of the time of day, in the relevant sense. One might demand a *strong* correlation by some quantitative measure, but no measure of statistical correspondence will specify the semantic content that representations have [39,40].

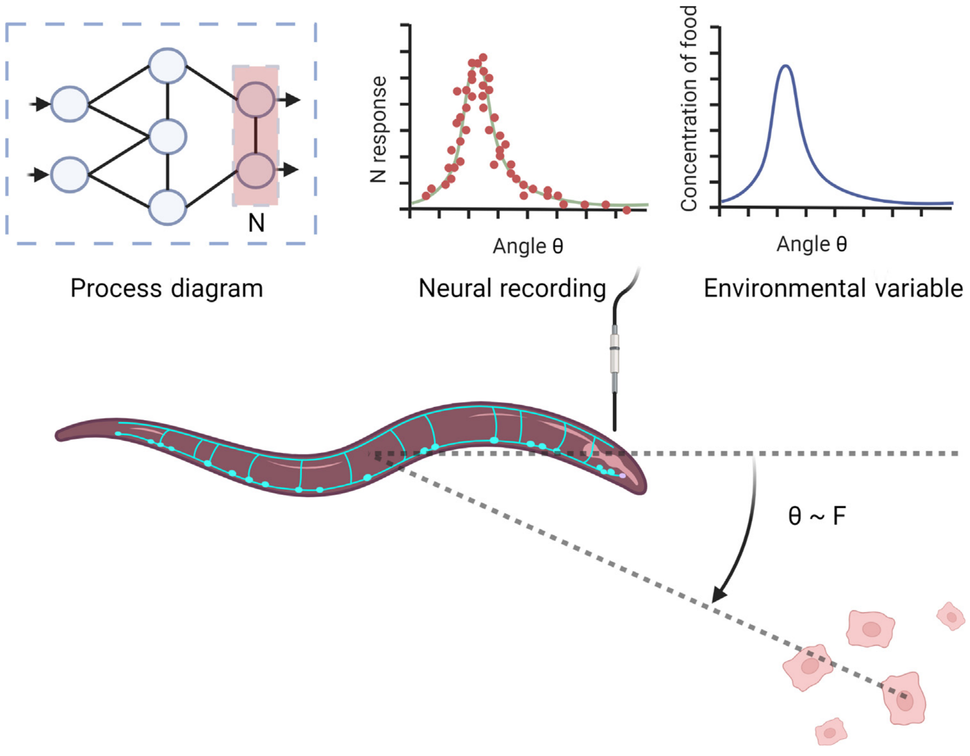
This is brought out most clearly by cases where the behavioral stakes are high and the cost of a false negative is significantly different from that of a false positive. Consider, for example, that many animals respond with avoidance behaviors to relatively simple features of visual stimuli, such as a looming dark object [41,42]. In these cases, neural activity over short time-spans may reliably lead to avoidance behavior where these neural activities correspond with 'looming' described in strictly visual terms, essentially as an expanding light-occlusion. However, it is widely assumed in such cases that the represented content merits the avoidance behavior. That is, the neural activity is typically held to represent something one should avoid, something that poses a threat, which an expanding light-occlusion is not. A standard view of the content of a representation in such cases depends on something about the relationship between expanding light-occlusions and actual threats in an animal's environment. This is not captured by correlation strength, since the neural activity is more reliably correlated with (nonthreatening) expanding light-occlusions than with any relevant threat.

The notion that brain activity must in some way reflect the things being represented is well explored in neuroscience and might be called the predominant focus. Most systems neuroscience experiments correlate neural activity with features of an animal's environment (e.g., stimuli or movement) and draw general conclusions from these findings (Figure 1; we will examine chemotaxis in *Caenorhabditis elegans* in more depth later, using it as a central example). At least since the work of Edgar Adrian, the observed correspondence between neural activity and stimulus features was recognized as important and as potentially related to representation or coding. In point of fact, the Nobel in Physiology or Medicine was awarded in 1963 for discoveries about the mechanisms of electrical transmission allowing for signal communication in nerve cells [43] and in 2021 it was awarded for findings about the mechanisms of receptor cells involved in touch and temperature sensation (https://www.nobelprize.org/prizes/medicine/2021/press-release). Experimentally, such observations are possible even with single electrode recordings in conjunction with presented stimuli or observed behavior. Thus, correlational findings have formed the basis for many conceptions about how the brain represents features of the world and how the brain generates adaptive behavior.

Aspect 2: causal role

The second aspect follows from the way representations are supposed to help us to make sense of how an animal acts (i.e., from their role in causal explanations of behavior). Representations of features of the world that pose opportunities and threats are hypothesized as part of what drives creatures to act in regular ways: toward what they need, away from what is dangerous, etc. [9,44]. For example, a representation of a predator might be offered as part of a causal explanation of an animal's flight. Generally, this aspect says that *if neural activity N is a representation of F, then N's correlation with F is causally involved in some F-related behavior(s).*

Contemporary philosophical accounts widely assume or argue that representations figure in causal relations that mediate behavior. Various philosophers in the late 20th century were concerned with understanding the sort of causal connection that must exist between a representation, its contents, and other parts of a representational system. [16,25,27,45,46]. A lively discourse emerged around the issue, providing the setting for more recent philosophical approaches to representation [31,47–49]. All sides of this debate agree that correlation



Trends in Cognitive Sciences

Figure 1. Idealized depiction of correlation (Aspect 1). Shown here between neural activity in a worm and an environmental variable. The angle between continuing forward motion and the greatest concentration of nearby nutrients. A generic 'process diagram' in the upper left stands for a hypothetical model of the significant neural connections. This figure was created using BioRender (https://biorender.com/).

(Aspect 1) does not suffice to classify the kind of internal states of interest. Of course, neuroscientists and philosophers do not share all of the same explanatory interests, but here there seems plainly to be an overlap in the two fields' ideas about information-carrying states with a special, content-bearing role in adaptive behavior. Many philosophers have approached the relevant problem as being to answer what justifies calling some brain activity a representation of *F other than* (perhaps in addition to) the fact that the activity is found to correlate with *F* and many have pointed to some form of a causal role in behavior as part of the solution (Box 3).

Consider, for example, Soh *et al*.'s study concerning representations in the neural activity of the worm, *C. elegans* [50]. Simplifying slightly, the authors hypothesized that the direction of a higher concentration of nutrients (NaCl) was represented by the difference between the firing rates of two pairs of interneurons around the front of the creature's nose-tip. Note that some sensory neural activity synaptically upstream from the proposed representation transmits the information to those interneurons, which then affect the activity of other neurons, so the NaCl gradient correlates with many patterns of activity besides that of the two interneurons. In other words, the relevant information is present elsewhere and so, in principle, decodable from the activity of various structures other than the hypothesized representation by the activity of two specified neurons. Aspect 1 alone does not distinguish the interneuron activity from those other patterns of activity.

This second aspect can be specified further. Saying that neural activity *N* represents a feature *F* implies an effect on *F*-related behavior determined by the brain mechanisms that make use

Box 3. Mechanistic approaches to mind

Philosophers across quite distinct intellectual traditions have championed some version of the idea that even highly complex behaviors can be understood in terms of the coordination of internal physical mechanisms, described in terms of causal relationships. Descartes famously endeavored to rigorously specify how various behaviors are subserved by mechanisms in the nervous system [95]. Although many of Descartes' empirical hypotheses have been disconfirmed, and although he claimed certain 'higher' operations were carried out by a nonphysical mind (not subject to mechanistic analysis), his work exemplifies the basic aim of explaining behavior by appealing to neural mechanisms. Much longer ago, a broadly mechanistic view of complex behavior was also expressed in ancient India, in the Samkhya-Yoga school of thought. Samkhya divides the mind into three major components or organs, which are constituted by various kinds of internal physical changes and sensory-motor exchange with the world (see Patanjali's *Yoga Sutras*, especially the *Samkhya Karika* text and associated commentaries, available in [96]). According to Samkhya, the mind's three basic organs have the functional roles of: (i) registration of information and direction of movement (*manas*); (ii) integration of the ego [i.e., distinguishing self and mine from other and others' (*ahamkara*)]; and (iii) reflective thought and abstract reasoning (*buddhi*). There are arguably significant analogies between the Samkhya approach and contemporary western thinking about brain-based representational mechanisms [97]. The question of what defines a 'mechanism' or 'mechanistic explanation' has generated an influential and ongoing philosophical literature [98–101] and several contributors take special interest in mechanisms in the brain [102–105]. While our observational and theoretical tools have developed immensely in the modern era, historical philosophical approaches like those earlier exemplify the ongoing project of understanding cognitive capacities in terms of physically inner processes with specific causal relations to things in the outside world.

of *N* as a representation of *F*. The functional description of *N* as part of such a mechanism should support predictions about how experimentally blocking or inducing the representation should affect behavior. This makes *N* explanatorily relevant in a way that other activity that correlates with *F* is not.

In Soh *et al*.'s model of *C. elegans* chemotaxis, the representation of an NaCl gradient by two interneurons is described within a mechanism that tries to explain how a worm's salt-following and turning behaviors are generated. Such a mechanism incorporates a representation by showing how a state with some content (roughly, 'salt, *that way*') figures in complex processes that have some observable, behavioral output related to that content (Figure 2). Other parts of the *C. elegans* nervous system to which the two interneurons are closely related also shape the creature's following and turning behaviors. This illustrates the fact that various parts of a mechanism other than the representation of *F* also have effects on *F*-related behavior. Therefore, when investigating a mechanism hypothesized to describe how the brain might solve some problem, it is often a challenge to determine exactly which neural properties implement which parts of which mechanism. In creatures somewhat more complex than *C. elegans*, one also faces (more) challenging questions about how a mechanism is contingent on parts of the nervous system outside of wherever it is implemented. Thus, if some neural activity *N* correlates with a feature *F* and is found to have *some* effect on *F*-related behavior, *N* still might not be a representation of *F*. It could be that *N* represents something more general or specific than *F*, or *N* could be just a part of a larger representation of *F*, or *N* could be a causal relay involved in *F*-related behaviors without representing anything in particular.

Many patterns of neural activity involved in *F*-related behaviors are likely to modify a behavior in some fashion, but to say that all simultaneously play the role of representing *F* in the containing system would yield a deflated notion of 'representation'; it would imply an astonishing amount of redundancy and provide no significant explanatory relevance beyond correlation alone [51,52]. A specific causal role for *N* in generating behavior is part of what we mean in saying *N*, unlike other causally implicated neural patterns, is a representation of *F*. This means that, given a hypothesis as to what some neural activity represents, it will count against that hypothesis if neural activity with overlapping information and causal roles can be found in other brain activity, the role of which in the relevant behavior is not accounted for. In such a case the parallel stream

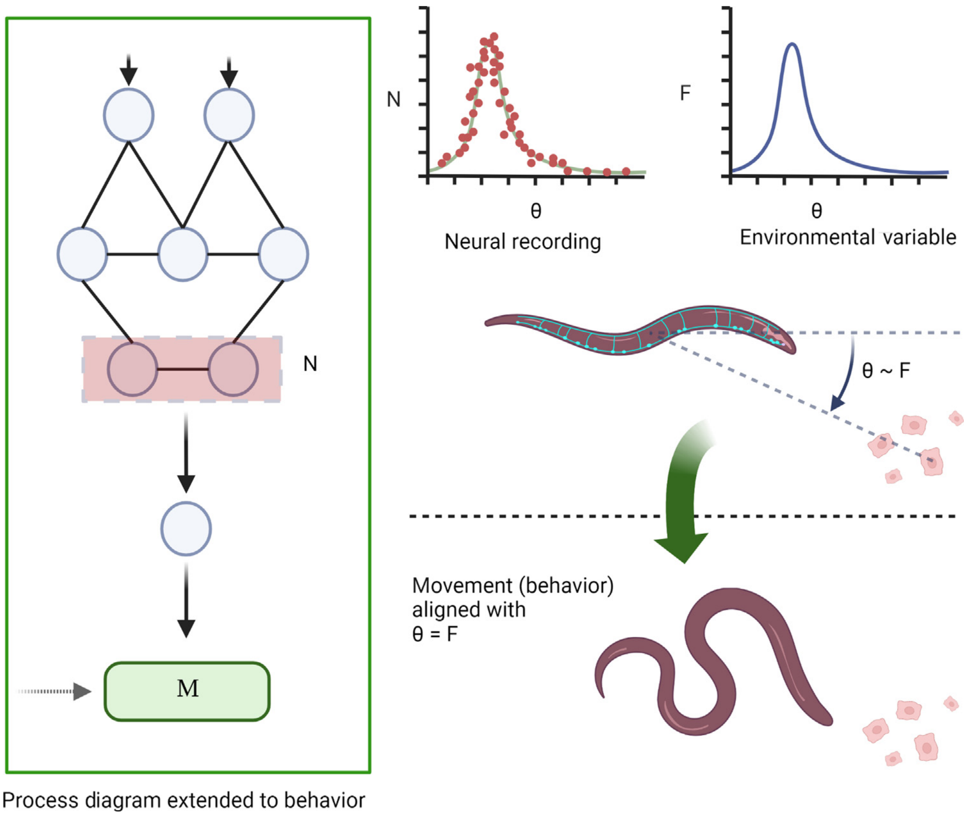


Figure 2. Idealized depiction of causal role (Aspect 2). Neural activity found to correlate with the direction of nutrients (from Figure 1) is shown to lead to a turning behavior in line with that direction, informing a mechanistic model of the relationship. This figure was created using BioRender (https://biorender.com/).

could indeed constitute a second representation of the same contents, but it could also be that one or both streams are causally involved but lack the distinctive causal role of a representation.

So what causal role(s) make for a representation? In Soh *et al*.'s model of *C. elegans*, for instance, the NaCl gradient information is relayed in some form through much of the nervous system, but some feature of how that information is processed between two interneurons must warrant locating the representation there; exactly what is this feature? Philosophers have tried in different ways to clarify the difference between mechanistic processes that merely happen to carry information and those that perform their causal role *by* carrying information. A particular point of contention has been whether mechanistic and representation-involving explanations include dynamic systems models of neural systems (and if so, which ones) [53–58]. This issue is further complicated by the fact that neuroscientists may be interested in representations at the level of an individual neuron, within a neural circuit, across brain areas, or even across animals. Resolving these issues in detail is beyond our present scope. All we want readers to take away here is the idea that neuroscience means to use representations in mechanistic explanations and, therefore, that representations are partly defined by their causal role in the behavior of a representation-using system.

In our estimation, this second aspect is usually not examined by systems neuroscience experiments, although it is by some, and establishing causal relations between neural activity and behavior may be increasingly feasible with the creative application of new tools [59]. While causality is often recognized as important to studying representations in neuroscience, establishing a causal effect on behavior for some neural circuits is more difficult and in many settings it is not well established [60]. Neuroscientists do focus on causal relations in certain cases, including in lesion and knockout experiments, electrophysiological stimulation [61,62], and optogenetics [63,64]. Such approaches aim to use controlled perturbation of neural activity, ideally noninvasively and with precision, to get at mechanisms that could relate the activity functionally to differences in behavior. Still, performing targeted interventions on the right neurons and only those neurons remains a challenge in many organisms. This makes it somewhat easier to make reasonably precise hypotheses about representations in a creature like *C. elegans*, whose entire nervous system activity throughout food-related locomotive behavior can be observed in detail, in real-time, and much of it intervened on at relatively low cost. Such complications notwithstanding, we find that this causal aspect is entailed in paradigmatic neuroscientific explanations involving representation (but see Box 4).

Aspect 3: teleology

The third aspect says that a representation must serve some aim, goal, or purpose ('telos' is ancient Greek for purpose). This aspect is highlighted by a consideration of cases involving mistakes or illusions, which deviate in a basic way from the more ordinary cases of successful perception and movement that we have considered so far. Such cases highlight the fallibility of representations, revealing an implied standard that actual representations either meet or fall short of. For a non-neural example, one normally says a map comes with a standard of accuracy; the fact that it is a map *of* (representing) a certain place implies a range of conditions that determine the extent to which it is a good map. Such accuracy conditions help explain, for instance, whether someone using the map can be expected to navigate successfully, or how they might be misled [64]. The map might correspond more or less to several different regions, but it helps explain the *mis*guided navigator's behavior only with respect to the region whose navigation it is supposed to serve.

Representations are supposed to allow neuroscientists to help explain patterns of misperception and misdirected behavior, as well as the more mundane 'good cases', and having a causal role in

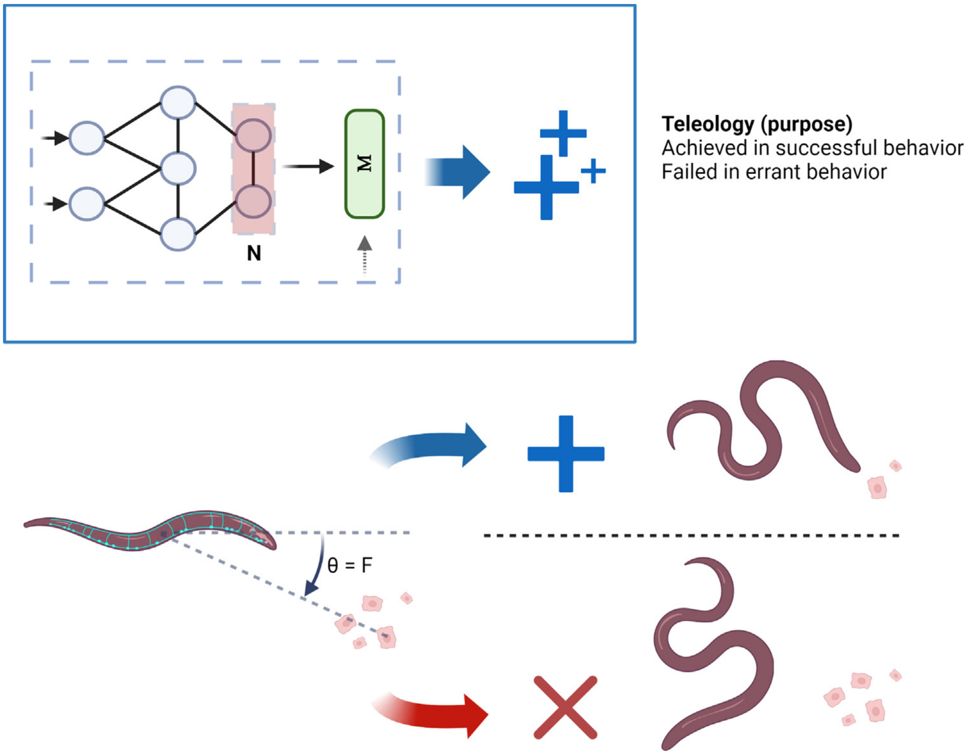
Box 4. A dissenting view

One can find descriptions of the relationship between representations and mechanisms in the brain that reject Aspect 2. A recent review article centers on 'brain-wide representations' in a few organisms, including *C. elegans*, suggesting that features of ongoing behavior are represented in much of their nervous systems [106]. The authors apply the term 'representation' to all neural activity that encodes features of ongoing behavior, suggesting that whether and how a representation is used is not essential to its status as such. On this approach, far more of the *C. elegans* nervous system than the two interneurons described in the research we discussed earlier would be counted as representing the worm's NaCl-tracking behavior. Similarly, some descriptions of place cells in the hippocampus suggest that they represent spatial features, regardless of whether they are used to interact with those spatial features. For instance, an overview by Hartley *et al*. [107] seems to imply that a correspondence relation (Aspect 1) suffices for representation, but notably their paper starts by asking how an animal is able to know where it is, remember distant goals, and navigate to them. A review of the topic by Colgin [108] highlights how many things *other than* spatial relations correspond to hippocampal activity in humans and relates these findings in terms of hippocampal 'representations' of various nonspatial aspects of experience, such as visual and temporal features. The terms of these authors give no indication that a representation of *F* implies a functional role for that representation. One philosopher who works on mechanistic explanation explicitly adopts this kind of view of place cell representations; in Bechtel's [109] terminology, the hypothesis about what hippocampal cells represent precedes any consideration of what the representations might be used for. While he does envision a mechanistic role in navigation for place cells, in his terms the content of these representations, the fact that they are about spatial features, does not depend on them playing such a role. Here we are arguing for a different way of framing things. It is not disputed that much of the *C. elegans* nervous system carries information about ongoing behavior and no mechanistic hypothesis is needed to see that the hippocampus carries information about a spatial layout. But that is just the first aspect. When it is further claimed that not just information about, but a 'representation of' food or place resides in some neural activity, this, we contend, rides on the assumption that they can play a functional role related to food- or place-related behavior.

behavior (Aspect 2) does not ensure this. Under a very strong evolutionary convergence assumption, one might expect every behavioral regularity (and the brain mechanisms it involves) to be adaptive and in this sense 'good', however it would be wrong to assume in general that evolution has finally converged. In any case, what teleology adds to the concept of representation is logically distinct from that of the causal role. Teleology allows a representation of *F* to tell us why an organism behaved *as if* it encountered an *F* even though it did not, reflecting the normative dimension of the way representations are commonly used in explanations in neuroscience. The telos or purpose needs to be specified so that *if N represents F, there is some basis for saying that N is* supposed to *correlate with F*, thereby establishing the possibility of misrepresentation [65,66]. This means that supposing one has identified a pattern of neural activity *N* whose correspondence with *F* (Aspect 1) plays a causal role in a process that modifies *F*-related behavior (Aspect 2), something further is needed to say why *N* represents *F* (in error) when the mechanism fails to support the behavior.

To illustrate this in the context of the earlier *C. elegans* example, we can first consider any teleological language used by the authors in describing the relevant processes. Soh *et al*. initially describe the movements involved in chemotaxis as strategies for survival and as being for 'approaching a favorable environment'. Perhaps we could aptly further specify the aim as 'approaching nutrients' (especially if there are favorable-but-nutrient-scarce environments), but either is fine here. We can imagine a mistake and a misrepresentation occurring in the following kind of case: part of the creature or environment outside of the proposed neural mechanism is affected so that the direction represented by the pair of interneurons is made to be *un*favorable. Suppose an extraneous electric current or an injury modifies the creature's sensory response so that the difference in the interneuron activity now leads the worm to swim 90 degrees to the right of where the highest concentration of NaCl is. If we have a clear enough view of the effect of the electrical current or injury, then the neural mechanism for chemotaxis and the included representation of the NaCl gradient can help explain the observed pattern of *mis*behavior. We would understand the worm to swim *as if* the salt were where it is heading, because of a (mis-) representation of the direction of the salt (Figure 3). So part of what one stands to gain in discovering how the creature's nervous system represents the NaCl gradient is an understanding of how it might behaviorally mistake an unfavorable environment for a favorable one (or a nutrient-scarce location for a nutrient-rich one). Since the content of the *C. elegans* representation is meant to contribute to our understanding of failure modes like this, we depend on the notion of a purpose or goal that the representation-containing process serves. More generally, fallibility with respect to some goal is essential to representations that neuroscientists try to identify. This is also brought out by research on illusions [67–69].

Arguably, a teleological view operates in the background of most neuroscientific research, implicit in the researchers' choices of which external variables to investigate and sometimes made explicit. For instance, in the Soh *et al*. study of chemotaxis in *C. elegans*, the authors assume that the relevant mechanism serves the purpose of survival, which might be tied to assumptions about its selection history (backward-looking) or tied to a view of the creature as one with the goal of surviving (forward-looking) (Box 5). For organisms like this, whose behavioral (and thus representational) repertoire is simple enough, one can be somewhat safe from doubt about the purpose some process serves. When there are few plausible purposes to choose from, backward- and forward-looking views will tend to align and be equally supported by widely shared knowledge. In such a case, carefully considering one's teleological assumptions makes little or no difference and the largely implicit or vague nature of appeals to teleology does not seriously hinder scientific communication about (representational and other) processes in the organism.



Trends in Cognitive Sciences

Figure 3. Idealized depiction of teleology (Aspect 3). The model of behavior (from Figure 2) is described within a purposive framework, supporting the characterization of turning behavior as achieving its purpose, or failing to. This figure was created using BioRender (https://biorender.com/).

However, when it comes to studying more complex behaviors with more contextual variability, it becomes nonobvious which goals an organism pursues and which mechanisms it relies on to do so and also nonobvious how such mechanisms may have served the reproductive successes of its ancestors (Box 6). For example, we might consider what is represented in the hippocampus of a rat or human. Supposing at first pass that hippocampal cells are used for navigation, we might say that place cell activity represents distances and directions to locations of expected reward in a known environment. However, one might also think the cells are used to represent spatial information in the course of *exploration*, which occurs before the environment is known, and therefore must involve representation of 'place' in some other sense. Further, there is evidence that hippocampal cells help support basic cognitive abilities often exercised in nonspatial domains, such as memory consolidation, planning, and imagination [70–72]. It seems fair to say that researchers are convinced that place cells are used by larger processes that allow for spatial navigation (perhaps among other capacities), but they do not entirely share a view of what those larger processes are. Studies in complex areas like this, therefore, tend to only provisionally or incompletely point to representations, as they assume the representations figure in still-to-be-established larger functional processes. Thus, with respect to representations of relatively abstract content in complex or unknown mechanisms, our conclusions should be somewhat more reserved and qualified so that we do not lose sight of the gaps in our knowledge we hope to fill.

An approach that brings teleology into the foreground in neuroscience is one that relies on data on the evolutionary development of the neural process under investigation. Some recent work in this spirit relies on knowledge about the phylogenetic tree to argue for constraints on the

Box 5. Two approaches to teleology

Philosophical accounts of the 'telos', or as it is sometimes called, the 'proper function' of some component, fall broadly into two kinds; backward-looking and forward-looking [25,31,46,110–114]. Backward-looking or 'etiological' approaches essentially hold that a component's proper function is what it did in the past to contribute to its current presence in the larger system. On this approach, a component's proper function explains 'why it is there' in a causal, evolutionary sense (i.e., its 'telos' is what it was naturally selected for). In the (organismal) systems of interest, the applicable notion of selection typically acts on reproductive lineages. For example, given generations of reproducing systems that all have a heart as a component, and assuming that it was past hearts' contribution to the circulation of blood that mainly explains the presence of similar hearts today, this view says that blood-circulating is the proper function of the heart. Hearts have other interesting characteristic properties that could be called functions, for instance, they make a rhythmic sound and reflect certain wavelengths of light more than others, but these do not qualify as the heart's proper function, on this approach. In analogous fashion, a backward-looking teleological view of neural representation would derive the content of such representations from the selection history of the neural activity.

The alternative is a forward-looking approach, sometimes called a goal-contribution view [115]. Forward-looking teleology is based on a prior definition of systems that exhibit goal-directed behavior, where the proper functions of components are determined by their contributions to specific goals in such systems. On this kind of view, one must start with some way of discerning which systems have goals, at least in general terms and perhaps also in formal terms. Often measures of mutability or robustness are taken to be evidence for the goal-directedness of some system's behavior. In essence, one examines how various are the circumstances from which the system will proceed to the outcome (goal) and how various are the perturbations or obstacles that the processes supporting that outcome can accommodate. Different philosophers have articulated these notions somewhat differently, but they similarly try to account for goals as outcomes that the system's behavior would adjust to ensure across a wide range of counterfactual scenarios [116–118].

This teleology at the system-level is what determines the proper roles of functional components; a component is supposed to do whatever contributes to the system's capacity to direct itself to the goal(s). This presents a forward-looking notion of 'why the component is there', which justifies treating other features of the component as functionally irrelevant and justifies treating some activities as malfunctional. On this view, for instance, one can claim that 'the heart is there to circulate blood' insofar as blood circulation is mainly how the heart contributes to some goal-directed activity of the bodily system in which it functions. On this approach, one might also claim that 'the heart (of a human) is there to make rhythmic sounds', if its production of these sounds supports the person's goals, say, by signaling to a doctor what must be done to preserve the person's life. Accordingly, some neural activity *N* would have the function of corresponding with some *F* only if doing so figured in an ensemble of functional components that together generate a goal-directed activity.

Whether just one of these two approaches to teleology can ultimately resolve all the cases of interest does not matter here (Box 6) and perhaps a pluralistic view will prove best [119]. Both views provide a frame for analyzing the components of living things that affirms the teleology that distinguishes them from other natural phenomena [120].

kinds of brain functions that neuroscience should study [73]. Although this work does not speak directly to identifying neural representations, it aligns with the general principle of looking to teleology derived from reproductive success to refine neuroscientific research. Another approach that foregrounds teleology starts by inferring organismal goals from macroscopic behavior and carefully observes the kinds of errors the organism is prone to make, looking for representations that seem essential to explain the observed patterns of successful and errant behavior [74,75]. Another approach is based on the teleology inherent in tasks used to train artificial neural networks; one optimizes for a particular task or set of tasks and compares the resulting network structure and activity to that of the brain, the idea being that how the learning goal shapes the developed neural system is essential to the workings of both artificial and biological cases [76–79].

An avenue of neuroscience concerned primarily with the third aspect is one that derives claims about the brain from a notion of coding efficiency, for instance, for neural systems in visual cortex [80,81]. A central hypothesis in this research is that, supposing animals have a visual world, the important common properties or events of which occur sparsely, it makes sense for them to have a visual system that uses a sparse coding scheme. This idea is supported by the finding that training a learning algorithm to represent natural scenes with sparse connections produces components with similar receptive fields as we see in visual cortex [82]. This explanatory

Box 6. Philosophical questions about teleology

A challenge for philosophers who favor backward-looking teleology is to precisely define the relevant notion of 'selection'. In doing so, one must address the so-called 'problem of novel contents', which means accounting for proper functions involving objects too novel to have influenced reproductive selection. For instance, parts of a human's brain might represent technologies or social structures that did not exist even one or a few generations ago. Relatedly, sometimes components seem to suddenly acquire new functions, as might have happened the first time wings were used for digging or hands used for signing. Defenders of a backward-looking approach have proposed ways of addressing these issues, which are beyond our scope here. Briefly, some have appealed to the idea of a component having a function to acquire a function, such that acquired functions can be shaped according to novel circumstances [110,121]. Garson [114] offers an account of selection in terms of differential retention in a population, which can occur without reproductive lineages and which, he argues, applies to the selection of synapses.

A forward-looking teleological view must meet the challenge of defining goal-orientedness in a way that is both measurable and appropriately restrictive [i.e., not assigning goals to any entities that we pre-theoretically agree have none, or denying goal-orientedness to entities that we are quite sure have goals (although some revision of pre-theoretic notions may be necessary)] [115]. Further, one must deal with the complicating possibility of one system having conflicting goals, or goals that change over time. There are forward-looking views with minimal requirements on the complexity of goal-directedness, for which systems like a prokaryotic cell, a Watt governor, and Ashby's homeostat [122], would all count as goal-directed [117]. Others defend more substantive requirements, typically including mechanisms to maintain a constantly decaying body that contains those mechanisms [116,118].

approach starts with a norm of efficiency tied to the goals and resources of a seeing animal and looks at how this norm may shape neural processing and architecture.

Neuroscientists' common aim of understanding the function of parts of the brain implies a basic commitment to some form of teleology, insofar as talk of 'functions' entails some notion of what fulfills the goals of a species or individual. However, teleological ideas are often hidden or ignored, as experimental data itself can usually only speak to the question of encoding or even only to correlation. Such a loose approach to representations provides no principled basis on which to distinguish the cases of misrepresentation, wherein the observed encoding relationship breaks down but the neurons are used as they normally are, leading to inappropriate behavior. To carefully investigate representations, observations of what neural activity correlates with should be integrated into a well-described causal and teleological framework. We hope that by clarifying how claims about neural representations stand with respect to these three aspects of representation, neuroscientists can communicate and test more precise and informative hypotheses, advancing the pace of learning about how neural systems shape the behavior of living things.

Concluding remarks

To close, we can briefly illustrate how the three aspects we distinguish apply to a pair of simplified examples. First, consider motion representation in human perception. The three-aspect definition offered here suggests the identification of such a representation in neural activity should be supported by: (i) finding that the neural activity correlates to (some property of) motion; (ii) intervention on the activity that affects motion perception-informed behaviors; and (iii) one or more hypotheses about how the motion perception enabled by the neural activity serves a (backward- or forward-looking) purpose. Second, consider how one might identify a visual representation of a friend's face: one could show that it: (i) correlates with that friend's face being within the field of view; (ii) plays a causal role in one's ability to pick one's friend out of a crowd; and (iii) that visually recognizing friends' faces (or those of social relations more broadly) serves one or more important purposes for the system in question. In general, one could establish any one or two out of these three aspects with respect to some neural activity. It follows from our view that, to the extent that it is doubtful that any one of these aspects can be established, one should question the hypothesis that the neural activity is a representation. Of course, there will be degrees of uncertainty and complication with each aspect; a variety of measures of correlation, causal interventions, and

Outstanding questions

What kinds of causal roles suggest that some neural activity bears representational content, rather than performing some nonrepresentational function?

What are the best experimental methods for distinguishing the different causal roles that a given pattern of neural activity might play?

What kinds of teleological frameworks are apt for neuroscientific inquiry and how does this vary (say, by subfield, scientist, or scale)?

What are good ways of quantifying and measuring performance (as successful or errant) according to these different teleological frameworks?

teleological premises to choose from, with different benefits and drawbacks (see Outstanding questions). Bringing in all three aspects implies that claims about representations should be more limited or contingent than they sometimes have been, but it allows for the communication of important parts of what we do and do not know about how some nervous system works and it reflects the ambitious explanatory task that the concept of representation is used for in neuroscience.

We have also called on neuroscientific research to take seriously the philosophical assumptions that underlie it. Philosophers are not typically well-positioned to make specific hypotheses about how neural systems work, but they have thought long and hard about the intricate conceptual interplay among notions of correlation, mechanism, teleology, behavior, explanation, and representation, among other ideas in terms of which to better understand the complex phenomena of intelligent life. Here we have offered an investigation of how neural systems research implicates some of these ideas, especially representation, grounded in a broad range of philosophical perspectives on the topic. We hope working to integrate these perspectives across disciplines serves the intellectual interests of neuroscientists, philosophers, and others.

Declaration of interests

No interests are declared.

References

1. Brette, R. (2018) Is coding a relevant metaphor for the brain? *Behav. Brain Sci.* 42, e215
2. Kragel, P.A. *et al.* (2018) Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron* 99, 257–273
3. Jones, I.S. and Kording, K.P. (2019) Quantifying the role of neurons for behavior is a mediation question. *Behav. Brain Sci.* 42, e233
4. Ritchie, J.B. *et al.* (2019) Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *Br. J. Philos. Sci.* 70, 581–607
5. Mirski, R. and Bickhard, M.H. (2019) Encodingism is not just a bad metaphor. *Behav. Brain Sci.* 42, e237
6. Vilarroya, O. (2017) Neural representation. A survey-based analysis of the notion. *Front. Psychol.* 8, 1458
7. Harvey, I. (2008) Misrepresentations. In *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, MIT Press
8. Yarkoni, T. (2022) The generalizability crisis. *Behav. Brain Sci.* 45, e1
9. Laplane, L. *et al.* (2019) Opinion: why science needs philosophy. *Proc. Natl. Acad. Sci. U. S. A.* 116, 3948–3952
10. deCharms, R.C. and Zador, A. (2000) Neural representation and the cortical code. *Annu. Rev. Neurosci.* 23, 613–647
11. Bickhard, M.H. (2009) The interactivist model. *Synthese* 166, 547–591
12. Clark, A. and Toribio, J. (1994) Doing without representing? *Synthese* 101, 401–431
13. Krakauer, J.W. *et al.* (2017) Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490
14. Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, The MIT Press
15. Newell, A. *et al.* (1958) Elements of a theory of human problem solving. *Psychol. Rev.* 65, 157–166
16. Fodor, J.A. (1990) *A Theory of Content and Other Essays*, MIT Press
17. Haugeland, J. (1978) The nature and plausibility of cognitivism. *Behav. Brain Sci.* 1, 215–226
18. Chalmers, D.J. (1994) On implementing a computation. *Mind. Mach.* 4, 391–402
19. Piccinini, G. (2008) Computation without representation. *Philos. Stud.* 137, 205–241
20. Caston, V. (2019) Intentionality in ancient philosophy. In *The Stanford Encyclopedia of Philosophy, Stanford University*
21. Pitt, D. (2020) Mental representation. In *The Stanford Encyclopedia of Philosophy, Stanford University*
22. Dennett, D.C. (1987) *The Intentional Stance*, MIT Press
23. Field, H.H. (1978) Mental representation. *Erkenntnis* 13, 9–61
24. Fodor, J.A. (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press
25. Dretske, F. (1988) The explanatory role of content. In *Contents of Thought* (Grimm, R.H. and Merrill, D.D., eds), University of Arizona Press
26. Hatfield, G. (1991) Representation in perception and cognition: connectionist affordances. In *Philosophy and Connectionist Theory* (Ramsey, W. *et al.*, eds), Psychology Press
27. Churchland, P.S. and Sejnowski, T.J. (1990) Neural representation and neural computation. *Philos. Perspect.* 4, 343–382
28. Markman, A.B. and Dietrich, E. (2000) In defense of representation. *Cogn. Psychol.* 40, 138–171
29. Grush, R. (2004) The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–396 discussion 396–442
30. Eliasmith, C. (2005) A new perspective on representational problems. *J. Cogn. Sci.* 6, 97–123
31. Shea, N. (2018) *Representation in Cognitive Science*, Oxford University Press
32. Millikan, R.G. (2020) Neuroscience and teleosemantics. *Synthese* 199, 2457–2465
33. Cao, R. (2012) A teleosemantic approach to information in the brain. *Biol. Philos.* 27, 49–71
34. Rupert, R.D. (2018) Representation and mental representation. *Philos. Explor.* 21, 204–225
35. Dretske, F. (1981) *Knowledge and the Flow of Information*, MIT Press
36. Bickhard, M.H. (2000) Information and representation in autonomous agents. *Cogn. Syst. Res.* 1, 65–75
37. Millikan, R.G. (2001) What has natural information to do with intentional representation? *R. Inst. Philos. Suppl.* 49, 105–125
38. Baker, B. (2021) Natural information, factivity and nomicity. *Biol. Philos.* 36, 26
39. Harman, G. (1983) Problems with probabilistic semantics. In *Developments in Semantics* (Orenstein, A. and Stern, R., eds), pp. 243–237, Haven Books
40. Hájek, A. (2007) The reference class problem is your problem too. *Synthese* 156, 563–585

41. Schiff, W. *et al.* (1962) Persistent fear responses in rhesus monkeys to the optical stimulus of "looming.". *Science* 136, 982–983
42. Zhao, X. *et al.* (2014) Visual cortex modulates the magnitude but not the selectivity of looming-evoked responses in the superior colliculus of awake mice. *Neuron* 84, 202–213
43. Huxley, A.F. (1963) The quantitative analysis of excitation and conduction in nerve. *Les Prix Nobel* 1963, 242–260
44. Gallistel, C. and King, A. (2009) *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Wiley-Blackwell
45. Horgan, T. (1989) Mental quausation. *Philos. Perspect.* 3, 47–76
46. Millikan, R.G. (1984) *Language, Thought, and Other Biological Categories: New Foundations for Realism*, MIT Press
47. Millikan, R.G. (2017) *Beyond Concepts: Unicepts, Language, and Natural Information*, Oxford University Press
48. Neander, K. (2017) *A Mark of the Mental: A Defence of Informational Teleosemantics*, MIT Press
49. Burge, T. (2010) *Origins of Objectivity*, Oxford University Press
50. Soh, Z. *et al.* (2018) A computational model of internal representations of chemical gradients in environments for chemotaxis of *Caenorhabditis elegans*. *Sci. Rep.* 8, 17190
51. Ramsey, W.M. (2007) *Representation Reconsidered*, Cambridge University Press
52. Egan, F. (2020) A deflationary account of mental representation. In *What are Mental Representations?* (Smortchkova, J. *et al.*, eds), Oxford Academic Press
53. Van Gelder, T. (1995) What might cognition be, if not computation? *J. Philos.* 92, 345–381
54. Chemero, A. (2001) Dynamical explanation and mental representations. *Trends Cogn. Sci.* 5, 141–142
55. Kaplan, D.M. and Bechtel, W. (2011) Dynamical models: an alternative or complement to mechanistic explanations? *Top. Cogn. Sci.* 3, 438–444
56. Zednik, C. (2011) The nature of dynamical explanation. *Philos. Sci.* 78, 238–263
57. Silberstein, M. and Chemero, A. (2013) Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philos. Sci.* 80, 958–970
58. Beer, R.D. and Williams, P.L. (2015) Information processing and dynamics in minimally cognitive agents. *Cogn. Sci.* 39, 1–38
59. Urai, A.E. *et al.* (2021) Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* 25, 11–19
60. Marinescu, I.E. *et al.* (2018) Quasi-experimental causality in neuroscience and behavioural research. *Nat. Hum. Behav.* 2, 891–898
61. Becker, R.F. (1953) The cerebral cortex of man. By Wilder Penfield and Theodore Rasmussen. The Macmillan Company, New York, N.Y. 1950. 248 pp. *Am. J. Phys. Anthropol.* 11, 441–444
62. Penfield, W. *et al.* (1950) *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*, Macmillan
63. Deisseroth, K. *et al.* (2006) Next-generation optical technologies for illuminating genetically targeted brain circuits. *J. Neurosci.* 26, 10380–10386
64. Schellenberg, S. (2019) Accuracy conditions, functions, perceptual discrimination. *Analysis* 79, 739–754
65. Dretske, F.I. (1986) Misrepresentation. In *Belief: Form, Content, and Function* (Bogdan, R., ed.), pp. 17–36, Oxford University Press
66. Neander, K. (1995) Misrepresenting & malfunctioning. *Philos. Stud.* 79, 109–141
67. Firestone, C. and Scholl, B.J. (2016) Cognition does not affect perception: evaluating the evidence for "top-down" effects. *Behav. Brain Sci.* 39, e229
68. Benjamin, A.S. *et al.* Shared visual illusions between humans and artificial neural networks. In *Cognitive Computational Neuroscience 2019 Proceedings*, Berlin, Germany
69. Schlaffke, L. *et al.* (2015) The brain's dress code: how the dress allows to decode the neuronal pathway of an optical illusion. *Cortex* 73, 271–275
70. Epstein, R.A. *et al.* (2017) The cognitive map in humans: spatial navigation and beyond. *Nat. Neurosci.* 20, 1504–1513
71. Eichenbaum, H. (2017) Prefrontal-hippocampal interactions in episodic memory. *Nat. Rev. Neurosci.* 18, 547–558
72. Ólafsdóttir, H.F. *et al.* (2018) The role of hippocampal replay in memory and planning. *Curr. Biol.* 28, R37–R50
73. Cisek, P. (2019) Resynthesizing behavior through phylogenetic refinement. *Atten. Percept. Psychophysiol.* 81, 2265–2287
74. Ng, A.Y. *et al.* (2000) Algorithms for inverse reinforcement learning. In *ICML Proceedings of the Seventeenth International Conference on Machine Learning*
75. Körding, K. (2007) Decision theory: what "should" the nervous system do? *Science* 318, 606–610
76. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu Rev Vis Sci* 1, 417–446
77. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365
78. Richards, B.A. *et al.* (2019) A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770
79. Yang, G.R. *et al.* (2019) Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297–306
80. Olshausen, B.A. and Field, D.J. (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325
81. Simoncelli, E.P. and Olshausen, B.A. (2001) Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216
82. Zylberberg, J. *et al.* (2011) A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Comput. Biol.* 7, e1002250
83. Perkel, D.H. and Bullock, T.H. (1968) Neural coding. *Neurosci. Res. Program Bull.* 6, 221–348
84. Rieke, F. (1997) *Spikes: Exploring the Neural Code*, MIT Press
85. Allen, W.E. *et al.* (2017) Global representations of goal-directed behavior in distinct cell types of mouse neocortex. *Neuron* 94, 891–907
86. Hirokawa, J. *et al.* (2019) Frontal cortex neuron types categorically encode single decision variables. *Nature* 576, 446–451
87. Shadlen, M.N. and Newsome, W.T. (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* 86, 1916–1936
88. Pillow, J.W. *et al.* (2008) Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454, 995–999
89. Zhao, X. and Kording, K.P. (2018) Rate fluctuations not steps dominate LIP activity during decision-making. *bioRxiv* Published online May 4, 2018. https://doi.org/10.1101/249672
90. Latimer, K.W. *et al.* (2015) Neuronal modeling. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349, 184–187
91. Hubel, D.H. and Wiesel, T.N. (1959) Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591
92. Hume, D. (2019) *A Treatise of Human Nature*, Oxford University Press
93. Locke, J. (1689) *An Essay Concerning Human Understanding*, Oxford University Press
94. Taieb, H. (2018) *Relational Intentionality: Brentano and the Aristotelian Tradition*, Springer
95. Descartes, R. (1972) *Treatise of Man: French Text with Translation and Commentary, Trans. Thomas Steele Hall*, Harvard University Press
96. Radhakrishnan, S. and Moore, C.A. (1957) *A Sourcebook in Indian Philosophy*, Princeton University Press
97. Perrett, R.W. (2001) Computationality, mind and value: the case of Sāmkhya-Yoga. *Asian Philosophy* 11, 5–14
98. Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*, Cambridge University Press
99. Machamer, P. *et al.* (2000) Thinking about mechanisms. *Philos. Sci.* 67, 1–25
100. Bechtel, W. and Abrahamsen, A. (2005) Explanation: a mechanist alternative. *Stud. Hist. Phil. Biol. Biomed. Sci.* 36, 421–441
101. Andersen, H.K. (2011) Mechanisms, laws, and regularities. *Philos. Sci.* 78, 325–331
102. Hardcastle, V.G. and Stewart, C.M. (2002) What do brain data really show? *Philos. Sci.* 69, S72–S82
103. Craver, C.F. (2007) *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, Clarendon Press

104. Boone, W. and Piccinini, G. (2016) The cognitive neuroscience revolution. *Synthese* 193, 1509–1534
105. Ryder, D. (2004) SINBAD neurosemantics: a theory of mental representation. *Mind Lang.* 19, 211–240
106. Kaplan, H.S. and Zimmer, M. (2020) Brain-wide representations of ongoing behavior: a universal principle? *Curr. Opin. Neurobiol.* 64, 60–69
107. Hartley, T. *et al.* (2014) Space in the brain: how the hippocampal formation supports spatial cognition. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 369, 20120510
108. Colgin, L.L. (2020) Five decades of hippocampal place cells and EEG rhythms in behaving rats. *J. Neurosci.* 40, 54–60
109. Bechtel, W. (2016) Investigating neural representations: the tale of place cells. *Synthese* 193, 1287–1321
110. Millikan, R.G. (1989) In defense of proper functions. *Philos. Sci.* 56, 288–302
111. Papineau, D. (1987) *Representation and Reality*, Blackwell
112. Wright, L. (1973) Function. *Philos. Rev.* 82, 139–168
113. Godfrey-Smith, P. (1994) A modern history theory of functions. *Noûs* 28, 344–362
114. Garson, J. (2017) A generalized selected effects theory of function. *Philos. Sci.* 84, 523–543
115. Nagel, E. (1977) Goal-directed processes in biology. *J. Philos.* 74, 261–279
116. Bigelow, J. and Pargetter, R. (1987) Function. *J. Philos.* 84, 181–196
117. Boorse, C. (2002) A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology* (Ariew, A. *et al*., eds), Oxford University Press
118. Mossio, M. *et al.* (2009) An organizational account of biological functions. *Br. J. Philos. Sci.* 60, 813–841
119. Preston, B. (1998) Why is a wing like a spoon? A pluralist theory of function. *J. Philos.* 95, 215–254
120. Mayr, E. *et al.* (2004) *What Makes Biology Unique?: Considerations on the Autonomy of a Scientific Discipline*, Cambridge University Press
121. Garson, J. and Papineau, D. (2019) Teleosemantics, selection and novel contents. *Biol. Philos.* 34, 36
122. Ashby, W.R. (1960) The Homeostat. In *Design for a Brain: The Origin of Adaptive Behaviour* (Ashby, W.R., ed.), pp. 100–121, Springer