Minds & Machines
DOI 10.1007/s11023-017-9437-2

Towards a Cognitive Neuroscience of Intentionality

Alex Morgan[1] · **Gualtiero Piccinini**[2]

Received: 20 January 2017 / Accepted: 19 May 2017

Abstract We situate the debate on intentionality within the rise of cognitive neuroscience and argue that cognitive neuroscience can explain intentionality. We discuss the explanatory significance of ascribing intentionality to representations. At first, we focus on views that attempt to render such ascriptions naturalistic by construing them in a deflationary or merely pragmatic way. We then contrast these views with staunchly realist views that attempt to naturalize intentionality by developing theories of content for representations in terms of information and biological function. We echo several other philosophers by arguing that these theories over-generalize unless they are constrained by a theory of the functional role of representational vehicles. This leads to a discussion of the functional roles of representations, and how representations might be realized in the brain. We argue that there's work to be done to identify a distinctively *mental* kind of representation. We close by sketching a way forward for the project of naturalizing intentionality. This will not be achieved simply by *ascribing* the content of mental states to generic neural representations, but by identifying *specific* neural representations that explain the puzzling intentional properties of mental states.

Keywords Intentionality · Representation · Cognitive neuroscience · Mechanisms

✉ Gualtiero Piccinini
piccininig@umsl.edu

[1] Rice University, Houston, TX, USA

[2] University of Missouri–St. Louis, St. Louis, MO, USA

Published online: 29 May 2017

1 Intentionality and the Rise of Cognitive Neuroscience

We *believe*, *desire*, *fear* things—these are among our *intentional* mental states. Intentional mental states are directed at things, such as flowers, fields, and fairies. Insofar as minds are capable of intentional mental states, intentionality is the mind being directed at things, which may or may not exist (Brentano 1874). This seems innocent enough, but on reflection we might wonder: How *could* the mind bear any relation to something, if that something doesn't exist? Ordinary physical relations are not like intentionality; the things they relate actually exist.

Yet intentional mental states also causally explain behavior, seemingly in virtue of what they're directed at. For example, Suzy's desire to meet Elvis explains why she went to Graceland; she went because of the specific Elvis-directed causal powers of her desire. Or so it seems. This is a central aspect of the puzzle of intentionality: intentionality seems to be a feature of the natural world, something that causally explains behavior, yet it relates one real thing—a mind—to something else that need not exist. This seems strikingly unlike anything else in nature.

In addition to having intentional mental states, our mind-brains contain and process internal *representations* such as perceptions, motor commands, mental images, mental models, and more. Such representations carry *information* about our internal and external environments. Or so cognitive scientists and neuroscientists tell us. Scientists explain our cognitive capacities in terms of specific computations over such representations—in terms of *information processing*. One foundational question in cognitive science and neuroscience concerns what it means for something to be an internal representation as well as for it to carry information, and how these notions are related.

Prima facie, the representations posited by cognitive scientists and neuroscientists should help solve the puzzle of intentionality. According to the Representational Theory of Intentionality (RTI), the intentionality of mental states can be explained by identifying mental states with the possession by a mind of appropriate representations. For example, the belief *that the light is on* is explained by a representation that stands in for the fact that the light is on and allows us to act accordingly. Similarly, the desire *that the light be on* is explained by a representation that stands for the same fact—that the light is on—but plays a different role: to motivate us to get the light to be on. In recent decades, RTI has become the mainstream view of intentionality (Jacob 2014; Pitt 2017).

Two features of RTI are worth pointing out from the outset. First, RTI privileges the intentionality of mental states over the intentionality of language. Linguistic expressions mean something, and meaning is often seen as a form of intentionality. This raises the question of whether linguistic or mental intentionality is more fundamental. We adopt the mainstream RTI according to which mental intentionality is more fundamental. Second, there is a long tradition of philosophers arguing that intentionality depends on phenomenal consciousness, so consciousness must be taken into account when explaining intentionality (Horgan and Tienson 2002; Loar 2003; Kriegel 2013; Bourget and Mendelovici 2017). By contrast, the mainstream

RTI theory that we adopt sidesteps consciousness altogether. The relation between intentionality and consciousness falls outside the scope of this paper.

In this opinionated review article, we argue that intentionality can be explained by cognitive neuroscience. We get there by tracing the historical development of RTI, addressing some of the central conceptual difficulties with the view, and discussing its prospects in light of the rise of contemporary cognitive neuroscience. Our broader goal is to shed light on the relation between intentional states and neural representations.

We begin in the next section by describing both intentionality and its relationship with representations more precisely. In Sect. 3, we discuss the explanatory significance of ascribing intentionality to representations, focusing on views that attempt to render such ascriptions naturalistic by construing them in a deflationary or merely pragmatic way. We then contrast these views, in Sect. 4, with staunchly realist versions of RTI, which attempt to naturalize intentionality by articulating conditions under which representations have a determinate content, expressed in terms of naturalistic notions such as information and biological function. We echo several others by arguing that these theories over-generalize unless they are constrained by a theory of the functional role of representational vehicles. This leads to a discussion in Sect. 5 of the functional roles of representations, and how representations might be realized in the brain. We respond to recent claims that the neurocomputational mechanisms posited by cognitive neuroscientists are not genuinely representational.

We argue that there's a clear sense in which they *are*, but that representations in this sense are also found in mindless systems, such as plants. Thus, there's work to be done to identify a distinctively *mental* kind of representation. We close in Sect. 6 by sketching a way forward for RTI: intentionality will not be explained simply by *ascribing* the content of mental states to generic neural representations, but by identifying *specific* neural representations, manipulated by appropriate neurocomputational mechanisms, which explain the puzzling intentional properties of mental states.

2 Intentionality and Mental Representation: A Historical Overview

Philosophers have discussed some of the puzzles associated with intentional phenomena since antiquity (Caston 2008), but the touchstone for modern discussions of intentionality is Franz Brentano's (1874) book *Psychology from an Empirical Standpoint*. Brentano characterized intentionality as the mind's directedness at "objects". Perceptual states or desires, for example, are *directed* at certain objects that are perceived or desired. Brentano held that this intentional directedness is the *mark of the mental*: it is exhibited by all and only mental states, so it demarcates the domain of psychology.

Brentano is widely interpreted as holding that the objects that intentional states are directed at may or may not exist in mind-independent reality. So in hallucinating an apple, one's mind is directed at something that doesn't really exist. This interpretation might not do justice to Brentano's considered view (Kriegel 2016),

but it has generated historically important discussions of the puzzles associated with intentionality. How could the mind be directed at "objects" that don't exist? How can we explain this puzzling phenomenon scientifically?

These puzzles have been extensively discussed in contemporary philosophy of mind, where intentionality is often characterized not in terms of Brentano's notion of directedness but in terms of *intentional content*, or the conditions under which a mental state is "satisfied" (Anscombe 1957; Searle 1983). Different kinds of mental state might be satisfied in different ways; for example, perceptions and beliefs are satisfied when they come to "fit" the world in the right way, whereas intentions or desires are satisfied when the world comes to "fit" them in the right way. Thus these mental states are said to have different "directions of fit". It's in virtue of being satisfiable in this way that a mental state counts as intentional, and the specific condition under which a mental state is satisfied is said to be (or be determined by) its intentional content. So, for example, the intentional content of Buffy's belief that Elvis is dead determines the condition under which the belief is *true*, namely the state of affairs in which Elvis is dead. Similarly, the content of Dizzy's desire to drink a martini determines the condition under which it is *fulfilled*, namely the state of affairs in which Dizzy drinks a martini.

How should we understand the relation between the intentional content and the directedness of a mental state? Much of the discussion of intentionality in the philosophy of mind has focused on mental states with a mind-to-world direction of fit, like perceptions or beliefs. Thus, as we'll see, the debate over naturalizing intentionality has focused on articulating naturalistic conditions under which intentional states correctly represent or *misrepresent* (e.g., Dretske 1986). From this perspective, it might be tempting to posit a simple relationship between intentional content and directedness: when a mental state is directed at an entity that doesn't exist, it misrepresents. But this needn't be the case; Buffy's belief that Elvis is dead is directed at something that no longer exists, namely Elvis, but it is nevertheless *correct*. Moreover, when we consider mental states with a world-to-mind direction of fit, the notion of misrepresentation is unhelpful. Dizzy's desire may be directed at a (non-existent) martini that confers eternal life, yet her desire doesn't misrepresent; it's just not the kind of thing that's evaluable in terms of (in)correctness.

Whether we think of intentionality as directedness or as intentional content, many have sought to resolve the puzzles of intentionality in terms of the manipulation of inner representations in the mind or brain. They've sought to develop a Representational Theory of Intentionality (RTI). This project appears already in the Early Modern philosophers, who appealed to the manipulation of mental representations—"ideas"—to explain, for example, how a subject can think of entities that she can't directly sense or perceive (e.g., Locke (1824 [1696]). Contemporary versions of RTI have taken their distinctive form as a result of certain developments in philosophy and psychology in the middle of the twentieth century.

Up until the early twentieth century, experimental psychologists sought to investigate mental phenomena in part through a methodology of careful, trained introspection. Starting with Watson's (1913) behaviorist manifesto, many psychologists repudiated this methodology as insufficiently rigorous and eschewed talk of unobservable entities such as representations. These *behaviorists* came to dominate

American psychology. They shifted the very subject matter of the science from mental processes to behavior.

Psychology shifted back towards mental processes in the middle of the century as a result of the so-called *cognitive revolution*, which was set in motion by the introduction of technical tools and concepts from communications engineering and computer science into psychology. Foremost among these were Turing's (1936) precise mathematical characterization of computation and Shannon's (1948) formalism for analyzing the efficiency of information transmission. These ideas influenced the development of the first rigorous computational theory of the mind-brain, proposed by McCulloch and Pitts (1943). They characterized neural networks in terms of what we'd now call logic gates, and they argued that psychological phenomena—including intentionality—could be explained in terms of computations carried out by circuits constructed from these networks (Piccinini 2004). With the emergence of artificial intelligence in the 1950s, this rich stock of ideas boiled over into the cognitive revolution, resulting in the displacement of behaviorism by the interdisciplinary field of cognitive science.

Cognitive science was founded on a computational theory of mind, which held that mental processes could be explained by appealing to the computational manipulation of inner, information-bearing representations. The neurally-inspired computationalism of McCulloch and Pitts (1943) soon gave way to an importantly different kind of computationalism, as cognitive science became increasingly influenced by symbolic artificial intelligence in the 1950s and '60s. The so-called *classical* computationalists held that computational theorizing about cognition should be pitched at a "functional" level of explanation that characterizes the *algorithms* involved in cognitive processing, while abstracting away from the neural mechanisms that implement those algorithms. This licensed a widespread assumption that cognitive theorizing is distinct and autonomous from neuroscience (Fodor 1975; Pylyshyn 1984).

Since the 1980s, however, cognitive theorizing has been increasingly constrained by mechanistic details about how the brain works, to the point that the mainstream paradigm in cognitive science is now cognitive *neuro*science. The computational explanations that classicists considered to be pitched at a distinct and autonomous functional level can be reinterpreted to be sketches of mechanisms, which might be gradually filled in with neuroscientific detail at an appropriate level of abstraction so as to eventually provide mechanistic explanations of the cognitive capacity of interest (Piccinini and Craver 2011). Contemporary cognitive neuroscience thus employs a computational and representational version of the explanatory strategy that is pervasive in the life sciences, whereby the capacities of complex organized systems are explained in terms of the parts and operations of mechanisms spanning multiple levels of organization (Bechtel 2008; Boone and Piccinini 2016).

Aside from its rejection of autonomy in favor of integrating neural and psychological levels, cognitive neuroscience is still recognizably continuous with the cognitive science that preceded it. What makes this tradition *cognitive*, it is widely thought, is that it harks back to the mentalism of pre-behaviorist introspectionist psychology—not by employing an introspectionist methodology, but by positing inner representations that have semantic content. Crucially, though,

the representations posited by cognitive scientists are accompanied by a rigorous theoretical framework that renders their causal powers non-mysterious. In this way, the notion of computation allows mental representations to overcome behaviorist qualms about "occult" mental entities, and find a home in experimental science.

How does all this relate to the nature of intentionality? To understand that, we need to look at a parallel stream of developments within philosophy, which eventually connect up to the cognitive sciences. Brentano's notion of intentionality as intentional directedness underwent significant changes at the hands of analytic philosophers, largely due to their methodological focus on the logical analysis of language. It was widely hoped that disputes about contentious phenomena could be resolved by analyzing the language used to talk about those phenomena, so attention shifted to sentences about mental states, rather than mental states themselves.

The most relevant application of this method was Chisholm's (1955, 1957) linguistic reformulation of Brentano's thesis, the view that intentionality is the mark of the mental. Chisholm held that intentional states are such that sentences describing them are *intensional*. Intensional sentences have three features: substituting co-referring terms (as in 'Hesperus' for 'Phosphorus') can change the sentence's truth-value, the presence of a singular term does not license existential generalization (because the putative referent of the singular term may not exist), and the whole sentence can be true even though the proposition that is the object of the intentional state is false (as when someone believes something false). Chisholm argued that intensional sentences could not be defined or paraphrased by sentences that did not contain mentalistic terms, such as sentences about behavior or other physical states of affairs. He concluded that intentional states cannot be redescribed in non-intentional terms.

Two other influences shaped conceptions of intentionality in contemporary philosophy of mind. The first was Quine's (1960) argument for the indeterminacy of meaning, according to which there is no objective inner fact of the matter that would resolve competing interpretations of what someone meant by a given sentence. The best interpretation(s) is just that which makes best sense of the person's behavior. This led Quine to a general skepticism about notions related to meaning, representation, and intentionality, and ultimately to the view that intentional states don't exist. Quine was a staunch physicalist, maintaining that only entities that are posited by our best scientific theories, which for him were physical theories, actually exist. He also accepted Chisholm's analysis to the effect that intentional states cannot be redescribed in non-intentional terms. Quine's conclusion was that, therefore, intentional states simply don't exist. This view initiates an influential *eliminativist* tradition about intentionality.

The second influence was Sellars's (1956) view that intentional mental states are unobservable entities hypothesized to exist by a tacitly endorsed theory, *folk psychology*. Unlike a scientific psychological theory, folk psychology isn't explicitly codified or experimentally tested. Nevertheless, folk psychology serves the same basic function of predicting and explaining behavior. This view came to be extremely influential in philosophy of mind, and much of the subsequent debate about the nature and existence of intentional states surrounded the relation between

folk psychology and scientific psychology. Unlike Chisholm and Quine, Sellars and many others held that the explanatory and predictive success of folk psychology provides strong reason to think that the theory is *true*, and that intentional states really *exist*. But how, precisely, do the intentional states posited by folk psychology relate to the computationally manipulable representations posited by cognitive science?

This allows us to bring together the two historical traditions we've been surveying and identify some of the central targets of our subsequent discussion. One highly influential version of RTI holds that in order for realism about intentional states to be vindicated, such states must be more or less *identified* with the representations of cognitive science. Proponents of this view allow that the kinds of intentional states recognized by folk psychology are probably too coarse-grained to pick out the actual causal determinants of behavior for serious scientific purposes. Nevertheless, they argue that realism would be vindicated by a science that is similar *enough* to folk psychology in that it posits causally efficacious inner states with intentional content—and, crucially, they hold that such a science is to be found in cognitive science. That is, proponents of this view hold that the representations posited by cognitive science have semantic content, and that this *just is* the intentional content of mental states.[1]

This strategy seemed to provide an appealing division of labor for solving the puzzles of intentionality from within the purview of natural science: Cognitive scientists would construct empirical theories of the mental representations that explain cognitive capacities, while philosophers would articulate a set of conditions, expressed in naturalistically respectable terms, that determine the semantic content of those representations. The hope was that this would allay Quinean concerns about indeterminacy and vindicate the idea that content plays an explanatory role in the etiology of behavior.

We'll spend much of the remainder of the paper evaluating the prospects for this project. But first we'll consider some alternative ways of understanding the explanatory significance of intentional ascriptions licensed by folk psychology, which treat such ascriptions as interpretation-dependent or merely pragmatic. This will provide us with a clearer view of the options for understanding the relation between intentional states and mental representations.

3 Interpretivism and Eliminativism About Intentionality and Representation

One way to develop Quine's skepticism about the intentional is to hold that the intentional states of a system are fundamentally a matter of interpretation. When a system behaves in sufficiently complicated and apparently rational ways, we may find it helpful to explain and predict its behavior as if it were controlled by intentional states, such as beliefs and desires. Crucially, the same behavior by the

[1] This strategy underlies a great deal of research in philosophy of mind throughout the late twentieth century; Fodor (1987) makes the strategy explicit.

same system might be interpreted in terms of different sets of intentional states. For example, we might find it compelling to explain why a mosquito bit me by attributing to the mosquito a desire to annoy me, plus the belief that I'll indeed be annoyed by its bite. That is an uncharitable interpretation. A more charitable interpretation is that the mosquito is looking for food and believes it can find it by biting me. But in principle there might be equally charitable and predictive interpretations of the behavior of a system.

Interpretivists about intentionality hold that, in case of multiple equally good interpretations, there is no deep fact of the matter that would determine which interpretation is uniquely correct (Davidson 1980; Dennett 1987). Multiple non-equivalent interpretations can be equally correct. Further, what makes an attribution of an intentional state to a system true is not that it refers to some concrete entity inside the system (e.g., in the mosquito's mind or brain), which exists independently of our interpretive acts. Rather, what makes such an attribution true is that it is licensed by the most charitable and predictive interpretation of the system's overall behavior. The ultimate arbiter of whether an interpretation is correct, for the interpretivist, is whether it is *useful*. Interpretivism is thus widely understood as a form of *instrumentalism*, according to which intentional states are simply a convenient fiction.

The same sort of interpretivism can be extended to the representations posited by cognitive scientists. The internal states posited by cognitive scientists are representations in the sense that such scientists interpret them as having semantic content. According to interpretivism, such content is not an objective property that representations have but one that is attributed by the interpreter. Again, the same internal representations may be interpreted in different ways, and there is no objective fact of the matter that decides which interpretation is correct (Dennett 1978; Cummins 1983, 1989; Churchland and Sejnowski 1992). Or perhaps the content ascribed to a system of representations is determined by the task being solved by the system; nevertheless, the content is a mere interpretative "gloss," ascribed merely for heuristic or expository purposes (Egan 2014).

Interpretivism is surely right in one respect: Ascribing semantic content to the computational processes putatively mediating cognition has proven heuristically useful in building models of cognitive processes. But this is a weak *epistemic* construal of the role of semantic content in cognitive science, which few would disagree with. Proponents of interpretivism about content have something stronger in mind; they hold that the computational processes mediating cognition don't objectively have semantic content independently of our interpretative practices.

What then are we to make of the apparent ubiquity and explanatory potency of ascribing semantic content to mental representations? A straightforward answer, more radical than interpretivism, is to simply deny that semantic content plays an important explanatory role in cognitive science. This is *semantic eliminativism*, which holds that although cognitive scientists do indeed traffic in entities that are often called "representations," these entities do not have semantic content; they are not representations *of* anything. When cognitive scientists posit "representations of X," they are not positing representations that stand in some representational relation to Xs; they are rather positing *X-type* representations, representations that are

individuated by their role in X-related cognitive processing. Semantic eliminativists allow that cognitive scientists might employ notions like *content*, *meaning*, or *reference* when presenting their theories, but they do so merely for expository convenience. These notions play no substantive explanatory role within the theories, and they may be abandoned entirely by more developed, mature theories (Stich 1983; Chomsky 1995).

If semantic eliminativism is correct, it remains unclear why ascriptions of content in cognitive science should be so useful and revealing. Semantic eliminativists hold that we're simply in *error* when we ascribe content to the computational mechanisms underlying cognition, but they fail to explain in a satisfying way why we're tempted to make that error. More generally, interpretivists and eliminativists about the role of content in cognitive theorizing have arguably done little to shed light on the explanatory relation between the intentionally-characterized cognitive processes that cognitive scientists seek to explain and the computational processes they posit as *explanantia*. If the states of computational mechanisms in the brain don't literally have content, how does positing such mechanisms shed light on intentional mental states? Simply put: how are we to explain intentionality?

This kind of worry has helped to motivate the version of RTI we mentioned earlier: the internal representations posited by cognitive science have semantic content, and this *just is* the intentional content of mental states. We'll now discuss the prospects for this view.

4 Tracking Theories of Intentionality and Representation

As mentioned at the end of Sect. 2, many philosophers in the late twentieth century hoped to solve the puzzles of intentionality by developing a naturalistic theory of intentional content, which would specify, in non-intentional terms, the conditions under which an intentional state has a determinate content. The hope was that a philosophical theory of content, together with a scientific theory of the causal powers of intentional states—presumably some mature cognitive science—would vindicate our ordinary practices of explaining behavior by appealing to the content of intentional states.

An early step in this direction was taken by Sellars (1954, 1956). As we've seen, Sellars argued that intentionality could find a place within a scientific theory of the world. In addition, Sellars construed intentional states as internal mental states analogous to sentences in a public language. Specifically, these internal mental states were causally related to one another as well as to environmental stimuli and responses analogously to how linguistic expressions are inferentially related to one another. Sellars proposed that the roles played by internal states within the cognitive economy of agents—mirroring the roles played by linguistic expressions within a language—also constitute their semantic content. This naturalistic account of intentionality, via the semantic content of internal states, later became known as functional role semantics (Harman 1970, 1973, 1988; Loar 1981; Block 1986; Peacocke 1992).

The main weakness of pure forms of functional role semantics is that the contents of many intentional states seem to depend fundamentally on the relation between those states and the world. To address this concern, proponents of so-called "long-armed" functional role semantics hold that part of an intentional state's functional role is its relation to the subject's environment (e.g. Harman 1987), but these proponents do little to elucidate the nature of this relation.

Other proponents of naturalizing intentionality looked at the relation between intentional states and the world as the starting point of their theories. They appealed to the way intentional states carry information about things as the basis for a naturalistic theory of content (Dretske 1981, 1988; Fodor 1987, 1990; Millikan 1984, 1993; Ryder 2004). After all, information clearly seems to be a semantic or representational phenomenon, and Shannon's (1948) communication theory delivers a precise and scientifically rigorous theory of information. However, it was widely recognized that communication theory provides a purely quantitative measure of the *amount* of information communicated by a signal, but doesn't determine the content of a specific signal. Nevertheless, it was also hoped that a causal interpretation of communication theory would help to identify the content of a signal, in roughly the following way: When *A* is reliably caused by *B* (in a way that satisfies the probabilistic constraints of communication theory), *A* is a signal with the information content that *B*. The resulting notion of information is more or less equivalent to what Grice (1957) called "natural meaning," what we have in mind when we say, for example, that smoke means fire.

But natural meaning is ubiquitous in nature, whereas intentionality presumably is not. Moreover, states that carry natural meaning aren't *ipso facto* evaluable as either correct or incorrect, yet it is widely held that in order for a state to genuinely represent anything, it has to be capable of *misrepresentation*. Proponents of information-based naturalistic theories of content sought to address these two concerns simultaneously by appealing to normative constraints on what a state is *supposed* to carry information about. The question of how to articulate such constraints in broadly naturalistic terms thus became the central agenda for the project of naturalizing semantics. Proposed answers were varied and ingenious, but they often appealed to a notion of biological *function*, or what a trait is *supposed* to do (Dretske 1988; Millikan 1984, 1993).

While this summary elides over a great many differences, proponents of naturalizing intentionality arrived at a shared conception of what a naturalistic theory of content should look like, and indeed of the very nature of intentionality. They conceptualized intentionality as a special kind of causal-informational relation between an internal state—a representation—and a distal entity, where the internal state has the function of responding to, or *tracking*, the distal entity. If the internal state *fails* to track that entity, it can be evaluated as *incorrect* or *inaccurate*. Theories in this broad family have been aptly dubbed "tracking" theories of intentional content (Tye 1995).

There was considerable controversy about whether the details of specific tracking theories can be worked out, or whether they deliver the intuitively correct content-assignments in specific cases. One central test case in the literature concerned the content of magnetosomes, organelles in certain marine bacteria that contain tiny

magnetic crystals. Magnetosomes help those bacteria reach their preferred conditions of anaerobic water (Dretske 1986). Debates raged about what the content of magnetosomes is—whether, for example, their function is tracking NORTH or ANAEROBIC WATER (e.g., Millikan 1989; Pietroski 1992; Jacob 1997). But all parties tacitly agreed that magnetosomes *have* content. This helps to underscore a problem about how tracking theorists understand the very nature of representation. As several philosophers have noted, even if biological functions can help to constrain the scope of tracking theories, such theories still encompass many informational states or structures—such as magnetosomes—that aren't representations in any explanatorily robust sense (Sterelny 1995; Ramsey 2007).

5 Representations in the Brain

The appeal to biological normativity allows tracking theorists to prevent many vehicles of natural information from counting as vehicles of intentional content, or representations. However, several philosophers have pointed out that tracking theories still encompass many states that don't seem to function as representations. For example, while the rings of a tree trunk presumably don't function to track the age of the tree and hence aren't representations, the magnetosomes mentioned earlier are widely thought to have the function of indicating the direction of anaerobic water, yet it is unclear whether magnetosomes should be regarded as genuine representations. They don't seem to "stand in" for anything—they just cause their bacterial host to move toward anaerobic conditions (Blakemore and Frankel 1981). Tracking theories thus seem to *over-generalize*. Some have argued that this is because tracking theories lack an adequate account of the functional role of *representational vehicles*: they tell us what the content of a given representation is, but not what makes something a representation in the first place (Ramsey 2007, 2016; Sterelny 1995). To vindicate intentionality, tracking theorists will have to say more about what counts as a representation.

Indeed, anyone seriously interested in the conceptual foundations of cognitive science must eventually grapple with what makes something a representation—perhaps most notably those who wish to argue that representations play no significant explanatory role in cognitive science. Since the early 1990s, several philosophers and cognitive scientists have developed arguments to this effect, typically drawing from various once-heterodox movements in cognitive science such as embodied cognition or dynamical systems theory (e.g., Brooks 1991; van Gelder 1995; Hutto and Myin 2013). Proponents of these arguments point to models in these new areas, which promised to explain cognitive phenomena without appealing to representations in any robust sense.

These arguments founder on two fronts. First, they tend to appeal to highly idealized models of relatively simple sensorimotor capacities, and it is unclear whether such models would continue to be non-representational when scaled-up to explain the kind of cognitive capacities for which representational explanations have seemed most urgent. As far as we can tell, none of the predictions from two decades ago of mature, non-representational explanations of distinctively cognitive

capacities have come to fruition. Notions related to embodiment and dynamical systems *have* proved fruitful—insofar as they have been taken up by the resolutely representational explanatory strategies of computational cognitive neuroscience. Second, and more problematically, such anti-representationalist arguments tend to target a stereotype of representations as static, word-like symbols, without clearly identifying fully general conditions for something to count as a representation. Representations might very well be action-oriented, dynamical, or both (Clark and Toribio 1994; Bechtel 1998; Grush 2003).

William Ramsey (2007) develops an anti-representationalist argument that avoids these worries. He pays close attention to the functional and explanatory roles in virtue of which something counts as a representational vehicle and argues that this role simply isn't occupied by the explanatory posits of the most mature and promising scientific explanations of cognition available—namely, those provided by cognitive neuroscience. Ramsey argues that the most fundamental condition for something to count as a representation is that it has semantic content, or *aboutness*. But, he argues, in order for ascriptions of content to be explanatorily robust, the vehicle of content must play a specific functional role that is *relevant to the content it has*.

To illustrate, consider a map of the NYC subway system. What makes this a representation is not merely that it is about the subway system in the sense that it reflects the abstract structure of the subway stops and their connectivity relations; after all, indefinitely many systems might adventitiously have this abstract structure. What makes the map a representation is that it is apt to be *used* to navigate the subway system *in virtue* of this abstract structural similarity. It serves as a *surrogate* for the subway system (Swoyer 1991).

This helps to illustrate a specific genus of representation that, Ramsey thinks, is especially significant in cognitive science: *structural representations*. Like models or maps, structural representations serve as surrogates for what they represent. Many philosophers and psychologists have developed structural notions of representation, often by appealing to the mathematical notion of a homomorphism, a structure-preserving map between two set-theoretic structures (Bartels 2006; Craik 1943; Cummins 1996; Isaac 2013; O'Brien and Opie 2004; Shepard and Chipman 1970).

An especially precise and detailed account of structural representation is provided by the psychologist Randy Gallistel (1990, 2008). For Gallistel, a structure A counts as a representation of a structure B just in case A is homomorphic with B, the homomorphism is causally mediated by a channel of information between A and B, and the manipulation of A allows the system of which A is a part to interact successfully with B. Homomorphisms in general are ubiquitous; but homomorphisms satisfying these conditions play a distinctive explanatory role—as Gallistel puts it, they are *functioning homomorphisms*.

Ramsey argues that structural representations were widely posited to explain cognitive capacities in the classical paradigm of cognitive science, but that classicism has largely been eclipsed in influence and explanatory power by cognitive neuroscience. There, Ramsey claims, we find little use for representations, structural or otherwise. Instead, we find mechanisms that are *called* representations, but which merely serve to detect certain entities in the environment. For example,

we see talk of "edge detectors" in the visual system, or "place cells" in the hippocampus that detect specific spatial locations.

Ramsey calls these mechanisms *receptors* and argues that although they function to detect distal entities, they don't serve as an internal *surrogate* for them; they essentially just function as triggers or causal relays, much like the sensors in automatic faucets. Ramsey holds that many philosophers have been inclined to think of receptors as representations because of the widespread tendency to treat tracking theories as theories of representational vehicles *as well as* theories of content. But many things function to *track* or *detect* without genuinely counting as representations—such as the magnetosomes that putatively represent the direction of anaerobic water. Thus, Ramsey argues, contemporary cognitive theorizing seems to be explaining cognitive capacities without appealing to representations in any robust sense.

Several philosophers have objected that Ramsey has an impoverished sense of the range and richness of the representational mechanisms on offer in cognitive neuroscience. In contrast to the simple feature detectors that Ramsey emphasizes, the mechanisms that have been the central focus of representational theorizing in neuroscience at least since Hebb (1949) are reverberating patterns of activity in populations of recurrently connected neurons. Contemporary computational neuroscientists analyze these patterns as dynamical attractors in a multidimensional phase space, the dimensions of which are determined by the activity levels of the neurons in the network that sustains the pattern of activity. Such a network is called an *attractor network* (Amit 1989). Depending on its connectivity and various other parameters, an attractor network might sustain attractors that trace various different manifolds, or "shapes," through its phase space; computational neuroscientists thus distinguish between point, line, ring, or even chaotic attractor networks, which are thought to play various different computational roles in the nervous system (Eliasmith and Anderson 2003).

Shagrir (2012) discusses an attractor network model of oculomotor control developed by Seung (1998) to illustrate that attractor networks generally function as structural representations. Seung's oculomotor network, thought to exist in the brainstem, settles into stable patterns of activity corresponding to points along a line attractor, each of which corresponds to a possible eye position. The set of possible states of the network, and relations between them, is *homomorphic* to the set of possible eye positions and their relations. Moreover, the current network state is hypothesized to encode a memory of eye position that serves to stabilize gaze. Thus the network is a *functioning* homomorphism in Gallistel's sense, hence qualifies as a model-like, structural representation. Indeed, Seung himself describes the network as an "internal model".

Similarly, Grush (2008) points out that an entire paradigm of neurocomputational theorizing about motor control, which employs mathematical tools from control theory, posits f*orward models* in the nervous system. These are hypothesized to encode the abstract structure of parts of the body or environment, to be updated via sensory information, and to run in parallel with motor processes to provide online feedback that enhances the speed and reliability of motor control. Forward models are also thought to be run *offline* in the service of cognitive processes such as

planning or counterfactual reasoning (Grush 2004). Again, forward models clearly qualify as structural representations. Indeed, forward models in the nervous system are widely thought to be implemented by attractor networks (e.g. Denève et al. 2007), and, conversely, the attractor network model of oculomotor control proposed by Seung (1998) can be understood as an implementation of a forward model.

Crucially, attractor networks and forward models are not merely exotica found only in far-flung nether regions of neuroscience; they play an indispensable role in the explanatory toolbox of modern computational neuroscientists, and they have been employed to shed light on a wide range of psychological phenomena, including motor control, sensorimotor integration, working memory, decision-making, mental imagery, and even social cognition (e.g., Rolls 2007; Wolpert et al. 2003; Wang 2001). *Contra* Ramsey, structural representations are central to the explanatory endeavors of mainstream cognitive neuroscience.

To underscore this, it's worth noting that even the feature detectors that Ramsey rests the bulk of his argument on are often characterized as representations by neuroscientists, not in virtue of their individual receptor functions, but in virtue of their role within a larger system that functions as a structural representation (Sprevak 2011). For example, the dominant neurocomputational models of hippocampal place cell function are attractor network models, which hypothesize that place cells contribute to a *cognitive map* of the spatial layout of an organism's environment.

According to a seminal model of this kind, developed by Samsonovich and McNaughton (1997), various sensorimotor signals about how far and in what direction an animal has moved drive a "bump" of activity within a two-dimensional plane attractor network comprised of place cells, such that the bump tracks the planar location of the animal within its current environment. This mechanism is patently a structural representation if anything is; a non-representational characterization of it would simply fail to elucidate how it contributes to an animal's capacity to navigate. Interestingly, Samsonovich and McNaughton are quite explicit that although place cells *contribute* to the representation, it is misleading to characterize them individually as representations. Indeed it seems plausible, more generally, that neuroscientists characterize feature detectors and other single cells as representations in virtue of their functional role within larger representational structures.

A potentially deeper objection to Ramsey simply dissolves his distinction between structural representations and mere receptors. Recall that Ramsey traces the tendency to think of receptors as representations to the influence of tracking theories of content. As he rightly points out, many tracking theorists pay scant attention to the functional and explanatory constraints on what counts as a representational vehicle of content. But not all do. In particular, Dretske (1988) pays close attention to questions about what representations are, and his tracking theory might reasonably be taken as a theory of content *and* representation. Ramsey thus allows that we might take Dretske's theory as canonical expression of the receptor notion, in much the same way we took Gallistel's theory as a canonical expression of the structural notion. Dretske holds, roughly, that a representation is a structure A that encodes natural information about some distal structure B and has been

selected through evolution or individual learning to guide certain activities with respect to *B*, in virtue of the fact that it encodes information about *B*.

As Morgan (2014) argues, however, the only substantive difference between Drestke's receptor theory of representation and Gallistel's structural theory is that the latter emphasizes that representing systems must be homomorphic to the systems they represent. But, Morgan argues, that's not a substantive difference at all, since it's a basic corollary of communication theory that if one system encodes information about another, a homomorphism holds between them. Thus anything that qualifies as a receptor *ipso facto* qualifies as a structural representation, and vice versa.

While systems that are intuitively receptor-like might not be prototypical of model- or map-like representations, they nevertheless stand in functioning homomorphisms with the systems they carry information about, so they qualify as structural representations on the clearest and most precise explication of that notion available, namely Gallistel's. Although these homomorphisms might be very simple, this surely doesn't preclude receptor-like systems from counting as representations; your car's oil light might represent that you're low on oil, even if it can only occupy two states. Whether a system counts as a representation doesn't depend on the number of *states* it can occupy, but—as Ramsey himself emphasizes—on whether the information content of those states is explanatorily relevant to what the system *does*. The theories put forward by Gallistel and Dretske clearly articulate this idea, so they seem to capture a perfectly legitimate notion of representation, a notion that subsumes even receptor-like systems.

From this perspective, we can see that tracking theories of intentionality and structural theories of representation converge on essentially the same basic phenomenon. While tracking theories are officially theories of content, they are motivated by a conception of the vehicles of content as informational states of representational systems *inside* the mind or brain. Tracking theorists often leave the nature of these vehicles unexplicated, but structural theories of representation fill this lacuna in part by emphasizing how the informational content of representational systems is relevant to the control functions of those systems.

While there are important differences about the details of tracking and structural theories to be worked out, these broad theoretical approaches are importantly complementary. They capture a clear, legitimate, naturalistic notion of representation that sheds light on explanatory practice in the cognitive sciences, notably cognitive neuroscience. Call this the *tracking* notion of representation. Importantly, this seems to be a sufficiently broad and ecumenical notion that it vindicates a wide variety of systems posited in cognitive (neuro)science. The interesting questions seem to be about what *kinds* of representations are needed to explain intentionality: what kind of structure and complexity must their homomorphisms have?

And there's the rub. Structural representations might not be as ubiquitous as bare vehicles of information, but they're still present in *all sorts* of natural systems, including *mindless* systems. For example, Morgan (2014) argues that circadian clocks in plants qualify as structural representations: They are homomorphic with the day-night cycle, they are entrained to the day-night cycle by information they receive from light signals, and they allow plants to perform certain activities that are

sensitive to the day-night cycle, such as reorienting their leaves to face the sun. The structural notion thus picks out a representational phenomenon that isn't distinctively mentalistic and does not seem to exhibit the kind of intentionality that minds have.

This isn't to suggest that there's something wrong with the notion. On the contrary, the notion is perfectly legitimate—it's the assumption that representational phenomena are inherently mentalistic that's problematic. Just as there is a tendency in the philosophical literature on intentionality to conflate a theory of content with a theory of representational vehicles, there is an equally pervasive tendency to suppose that a theory of representation provides a theory of *mental* representation. This is a mistake. We must look for something more specific than the notion of tracking representation for an account of what makes representations mental and gives rise to mental-grade intentionality.

6 Mental Representations and Mechanistic Explanation

We've now come full circle and are in a position to revisit our earlier discussion of intentionality. Insofar as tracking theories of intentionality encompass structures in mindless systems, such theories are in tension with a central aspect of Brentano's Thesis: That intentionality is an essentially mentalistic phenomenon. Indeed, many tracking theorists explicitly reject Brentano's Thesis (e.g., Fodor 1990; Millikan 2000). They do so because they take this to be the proper path towards naturalizing intentionality. Their guiding assumption is that if a phenomenon is distinctively mentalistic, then it is *irreducibly* mentalistic. So to show that intentionality is in fact reducible and continuous with the rest of nature, we must show that it is not *distinctively* mentalistic after all.

But reconceptualizing intentionality so as to encompass states of mindless systems leaves out one of its most important aspects. It's the distinctively mentalistic phenomenon of directedness towards entities that may not exist that poses the central puzzle of intentionality. That is not something that merely tracking representations, such as circadian clocks in plants, are able to do. Perhaps tracking representations contribute to *explaining* intentional directedness, but mental-grade intentional directedness seems to include something that is left unexplained by simply positing tracking representations.

Moreover, the tracking theorist's motivation for reconceptualizing intentionality as something that might be exhibited by mindless systems rests on a problematic conception of naturalization. It's a mistake to think that if intentionality is *distinctively* mentalistic, then it must be *irreducibly* so. Consider, by analogy, the case of life, which until relatively recently was widely regarded as difficult to accommodate within a naturalistic framework. Life is a distinctively organic phenomenon, but from that it doesn't follow that it is *irreducibly* organic. Biochemists and molecular biologists didn't naturalize life by expanding the boundaries of the concept so as to encompass inorganic phenomena such as rocks. Rather, they did so by identifying various capacities that are distinctive of living organisms and elucidating how those capacities are produced by mechanisms

nested across multiple levels of organization, which eventually bottom out in inorganic phenomena. That is, they didn't horizontally expand what counts as alive, but vertically elucidated the mechanistic basis of life. Life is *both* an organic phenomenon *and* explicable in terms of inorganic phenomena.[2]

Replace 'life' with 'intentionality', and 'organic' with 'mental', and we have a schema for how intentionality might be naturalized through the mechanistic explanations provided by contemporary computational cognitive neuroscience.[3] Importantly, naturalization so construed doesn't consist in providing a conceptual reduction of the kind that philosophers in the analytic tradition have often sought, but rather in providing a multilevel mechanistic explanation. This way of framing things can help guide future inquiry into solving the puzzles about intentionality we've been grappling with.

Existing attempts to elucidate the naturalistic basis of intentionality at best provide us with accounts of the semantic content of generic neural representations. Attempting to explain the intentional states of subjects by straightforwardly ascribing the intentional content of mental states to generic neural representation, whether we treat such ascriptions as literal or as a mere interpretative "gloss," assimilates generic neural representations to circadian clocks in plants. That is not an adequate theory of intentionality.

A more adequate theory of intentionality would recognize that tracking theories take us part of the way towards explaining intentionality, but not the whole way. What we also need is to identify *specific* neural representations, manipulated by appropriate neurocomputational mechanisms, which explain the puzzling intentional properties of mental states.

Consider one of the most puzzling aspects of intentionality: directedness towards non-existent "objects" such as centaurs. To explain it, we cannot simply posit neurons whose function is tracking centaurs. If there were such neurons, they would never fulfill their function. And since tracking theories assign content based on the function a representation performs, tracking theories would assign no content to such neurons. So positing neurons whose function is tracking nonexistent entities, in combination with tracking theories, is not going to explain our ability to think about nonexistent entities.

Yet when we think of a centaur, presumably some of our neurons are firing at a higher rate than when we do not think of a centaur. Given that they cannot simply get semantic content by *tracking* what they represent, we need a different theory to assign semantic content to them. One way they could get their content is by being composed of neural representations that do have tracking functions.

A neural representation of a centaur may be composed of a neural representation of human male head and torso, plus a neural representation of the body of a horse,

[2] For influential philosophical discussions of the mechanistic explanatory paradigm that is widely regarded as the dominant mode of explanation throughout the life sciences, see Bechtel and Richardson (2010), Machamer et al. (2000). For a discussion of how attempts to explain life fit within the mechanistic paradigm, see Bechtel (2011).

[3] For arguments that computational cognitive neuroscience provides psychological explanations that fit within the mechanistic paradigm, see Bechtel (2008), Kaplan (2011), Piccinini and Craver (2011), Boone and Piccinini (2016).

plus a neural representation of an appropriate geometric relation between the two, appropriately bound together through an appropriate binding mechanism. Even though the whole neural representation has no referents in the actual world, it still has representational content inherited from its constituents, each of which has real referents that can be tracked. The whole representation represents a centaur because a centaur is, roughly, a human male torso attached to the body of a horse. This sketch is simplistic, but it should convey the idea. A full version of this theory should specify which neural systems are involved in representing nonexistent entities and how they relate to systems that represent real things (for a start, see Piccinini, forthcoming).

A full neurocognitive explanation of intentionality will span mechanisms nested across multiple levels of organization in the nervous system and draw from ongoing developments in cognitive neuroscience. And there are other questions to keep philosophers busy: In virtue of what is mental content *mental*? Is mental content a distinct *kind* of content, or is it just the content had by mental states, which might in principle be had by other representational vehicles?[4] And, ultimately: How are mental representations integrated and coordinated such that their semantic content is intelligibly attributed to psychological subjects? The complementary efforts of philosophers and cognitive neuroscientists will shed naturalistic light on these questions.

Acknowledgements Thanks to Anne Jacobson, Jonathan Kaplan, Nelson Mauro Maldonado, Marcin Miłkowski, Alessio Plebe, Michael Schmitz, Leonardo Wykrota, and two anonymous referees for helpful comments. This material is partially based upon work supported by the National Science Foundation under Grant No. SES-1654982.

References

Amit, D. (1989). *Modelling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Anscombe, E. (1957). *Intention*. Ithaca, NY: Cornell University Press.

Bartels, A. (2006). Defending the structural concept of representation. *Theoria, 21*(55), 7–19.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22*(3), 295–318.

Bechtel, W. (2008). *Mental mechanisms*. New York: Taylor and Francis.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science, 78*(4), 533–557.

Bechtel, W., & Richardson, R. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press.

Blakemore, R., & Frankel, R. (1981). Magnetic navigation in bacteria. *Scientific American, 6,* 58–65.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10,* 615–678.

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese, 193*(5), 1509–1534.

Bourget, D., and Mendelovici, A. (2017). Phenomenal intentionality. In Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), forthcoming https://plato.stanford.edu/archives/spr2017/entries/phenomenal-intentionality/.

Brentano, F. (1874 [1995]). *Psychology from an empirical standpoint*. London: Routledge.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159.

[4] The distinction here echoes the distinction between 'content' and 'state' views of non-conceptual content (Heck 2000).

Caston, V. (2008). Intentionality in ancient philosophy. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition). http://plato.stanford.edu/archives/fall2008/entries/intentionality-ancient/.
Chisholm, R. (1955). Sentences about believing. *Proceedings of the Aristotelian Society, 56,* 125–148.
Chisholm, R. (1957). *Perceiving: A philosophical study*. Ithaca: Cornell University Press.
Chomsky, N. (1995). Language and nature. *Mind, 104,* 1–61.
Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.
Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*(3), 401–431.
Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
Cummins, R. (1983). *The nature of psychological explanation*. Cambridge: MIT Press.
Cummins, R. (1989). *Meaning and mental representation*. Cambridge: MIT Press.
Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.
Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.
Dennett, D. (1978). *Artificial intelligence as philosophy and as psychology. In his 'Brainstorms'* (pp. 109–126). Cambridge, MA: MIT Press.
Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Denève, S., Duhamel, J.-R., & Pouget, A. (2007). Optimal sensorimotor integration in recurrent cortical networks: A neural implementation of kalman filters. *Journal of Neuroscience, 27*(21), 5744–5756.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Form, content, and function* (pp. 157–173). Oxford: Clarendon.
Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170*(1), 115–135.
Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, Representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.
Fodor, J. (1975). *The language of thought*. New York, NY: Crowell.
Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.
Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
Gallistel, C. (1990). Representations in animal cognition: An introduction. *Cognition, 37*(1–2), 1–22.
Gallistel, C. (2008). Learning and representation. In J. Byrne (Ed.), *Learning and memory: A comprehensive reference* (pp. 227–242). Amsterdam: Elsevier.
Grice, P. (1957). Meaning. *Philosophical Review, 66,* 377–388.
Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy, 18*(1), 53–93.
Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*(3), 377–396.
Grush, R. (2008). *Representation reconsidered by William M. Ramsey*, Notre Dame Philosophical Reviews.
Harman, G. (1970). Sellars' semantics. *The Philosophical Review, 79*(3), 404–419.
Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Harman, G. (1987). (Non-solipsistic) conceptual role semantics. In E. Lepore (Ed.), *New directions in semantics*. London: Academic Press.
Harman, G. (1988). Wide functionalism. In S. Schiffer & S. Steele (Eds.), *Cognition and representation*. Boulder: Westview.
Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.
Heck, R. (2000). Nonconceptual content and the 'Space of Reasons'. *Philosophical Review, 109*(4), 483–523.
Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 520–933). Oxford: Oxford University Press.
Hutto, D., & Myin, E. (2013). *Radicalizing enactivism*. Cambridge, MA: MIT Press.
Isaac, A. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy, 91*(4), 683–704.
Jacob, P. (1997). *What minds can do: Intentionality in a non-intentional world*. Cambridge: Cambridge University Press.
Jacob, P., (2014). "Intentionality", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2014/entries/intentionality/.

Kaplan, D. (2011). Explanation and description in computational neuroscience. *Synthese, 183*(3), 339–373.
Kriegel, U. (Ed.). (2013). *Phenomenal intentionality*. Oxford: Oxford University Press.
Kriegel, U. (2016). Brentano's mature theory of intentionality. *Journal for the History of Analytical Philosophy, 4*(2), 1–15.
Loar, B. (1981). *Mind and meaning*. Cambridge: Cambridge University Press.
Loar, B. (2003). Phenomenal intentionality as the basis of mental content. In M. Hahn & B. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of Tyler Burge* (pp. 229–258). Cambridge, MA: MIT Press.
Locke, J. (1824 [1696]). *The works of John Locke* (Vol. 3, 12th Ed.). Rivington: London.
Machamer, P., et al. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115–133.
Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
Millikan, R. (1989). Biosemantics. *The Journal of Philosophy, 86*(6), 281–297.
Millikan, R. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.
Millikan, R. (2000). Naturalizing intentionality. *The Proceedings of the Twentieth World Congress of Philosophy, 9,* 83–90.
Morgan, A. (2014). Representations gone mental. *Synthese, 191*(2), 213–244.
O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezac (Eds.), *Representation in mind* (pp. 1–20). Amsterdam: Elsevier.
Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.
Piccinini, G. (2004). The first computational theory of mind and brain: A close look at McMulloch and Pitts's 'Calculus of Ideas Immanent in Nervous Activity'. *Synthese, 141*(2), 175–215.
Piccinini, G. (forthcoming). Nonnatural mental representation. In K. Dolega, T. Schlicht, J. Smortchkova (Eds.), *What are mental representations?* Oxford: Oxford University Press.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283–311.
Pietroski, P. (1992). Intentionality and teleological error. *Pacific Philosophical Quarterly, 73*(3), 267–282.
Pitt, D. (2017). "Mental Representation", *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.), forthcoming https://plato.stanford.edu/archives/spr2017/entries/mental-representation/.
Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.
Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas in Psychology, 40,* 3–12.
Rolls, E. (2007). *Memory, attention, and decision-making: A unifying computational neuroscience approach*. Oxford: Oxford University Press.
Ryder, R. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind and Language, 19*(2), 211–240.
Samsonovich, A., & McNaughton, B. (1997). Path Integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience, 17*(15), 5900–5920.
Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Sellars, W. (1954). Some reflections on language games. *Philosophy of Science, 21,* 204–228.
Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1*(19), 253–329.
Seung, S. (1998). Continuous attractors and oculomotor control. *Neural Networks, 11*(7–8), 1253–1258.
Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science, 63*(3), 519–545.
Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(7), 379–423.
Shepard, R., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology, 1*(1), 1–17.
Sprevak, M. (2011). Review of *Representation Reconsidered* by William Ramsey. *British Journal for the Philosophy of Science, 62,* 669–675.

Sterelny, K. (1995). Basic minds. *Philosophical Perspectives, 9,* 251–270.
Stich, S. (1983). *From folk psychology to cognitive science: The case against belief*. Cambridge, MA: MIT Press.
Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*(3), 449.
Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 42*(2), 230–265.
Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.
van Gelder, T. (1995). What might cognition be, if not computation. *The Journal of Philosophy, 92*(7), 345–381.
Wang, X. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences, 24*(8), 455–463.
Watson, J. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158.
Wolpert, D., et al. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1431), 593–602.