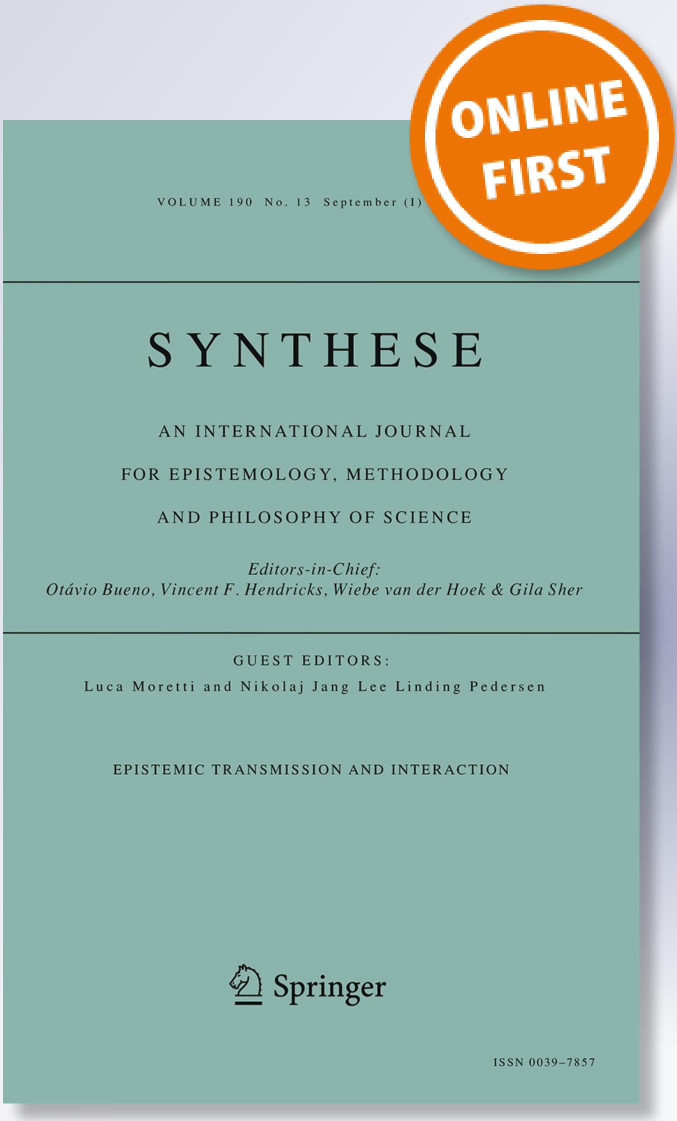
Representations gone mental

Alex Morgan

Synthese
An International Journal for Epistemology, Methodology and Philosophy of Science

ISSN 0039-7857

Synthese
DOI 10.1007/s11229-013-0328-7



ONLINE FIRST

VOLUME 190 No. 13 September (I)

SYNTHESE

AN INTERNATIONAL JOURNAL
FOR EPISTEMOLOGY, METHODOLOGY
AND PHILOSOPHY OF SCIENCE

Editors-in-Chief:
Otávio Bueno, Vincent F. Hendricks, Wiebe van der Hoek & Gila Sher

GUEST EDITORS:
Luca Moretti and Nikolaj Jang Lee Linding Pedersen

EPISTEMIC TRANSMISSION AND INTERACTION

Springer

ISSN 0039–7857

Springer

Synthese
DOI 10.1007/s11229-013-0328-7

Representations gone mental

Alex Morgan

Received: 14 January 2013 / Accepted: 26 July 2013

Abstract Many philosophers and psychologists have attempted to elucidate the nature of mental representation by appealing to notions like isomorphism or abstract structural resemblance. The 'structural representations' that these theorists champion are said to count as representations by virtue of functioning as internal *models* of distal systems. In his 2007 book, *Representation Reconsidered*, William Ramsey endorses the structural conception of mental representation, but uses it to develop a novel argument against representationalism, the widespread view that cognition essentially involves the manipulation of mental representations. Ramsey argues that although theories within the 'classical' tradition of cognitive science once posited structural representations, these theories are being superseded by newer theories, within the tradition of connectionism and cognitive neuroscience, which rarely if ever appeal to structural representations. Instead, these theories seem to be explaining cognition by invoking so-called 'receptor representations', which, Ramsey claims, aren't genuine representations at all—despite being *called* representations, these mechanisms function more as triggers or causal relays than as genuine stand-ins for distal systems. I argue that when the notions of structural and receptor representation are properly explicated, there turns out to be no distinction between them. There only *appears* to be a distinction between receptor and structural representations because the latter are tacitly conflated with the 'mental models' ostensibly involved in offline cognitive processes such as episodic memory and mental imagery. While structural representations might count as genuine representations, they aren't distinctively *mental* representations, for they can be found in all sorts of non-intentional systems such as plants. Thus to explain the kinds of offline cognitive capacities that have motivated talk of mental models, we

A. Morgan (✉)
Department of Philosophy, Rutgers University of New Jersey,
New Brunswick, NJ 08901, USA
e-mail: amorgan@philosophy.rutgers.edu

Published online: 29 August 2013

must develop richer conceptions of mental representation than those provided by the notions of structural and receptor representation.

Keywords Representation · Isomorphism · Psychological Explanation · Mental models · Neural networks · Feature detectors · Circadian clocks · Dretske · Eliminativism

1 Introduction

Behaviorists disparaged the introspection-based methodologies of earlier 'mentalistic' traditions in psychology as unscientific, and held that talk of internal mental states is explanatorily vacuous. The computer metaphor helped dislodge this attitude, and thereby set the cognitive revolution in motion, by providing a way of understanding how *something like* mental states—namely, symbolic representations—could play a causal and explanatory role within a purely mechanistic, scientifically explicable system. The cognitivist's appeal to representations betokened a return to mentalism, since although such representations were conceived of as 'sub-personal' and inaccessible to introspection, they were also regarded as *intentional* and somehow *internal* to the mind. While specific conceptions of computation and representation have changed with the fashions in cognitive science, a broad commitment to explaining cognitive capacities by appealing to the manipulation of mental representations has remained constant.

At least until recently. Since the mid-nineties, proponents of novel theoretical perspectives such as situated robotics and dynamical systems theory have argued that large swaths of cognition can be explained without appealing to representations at all. However, these approaches have tended to be most successful when explaining relatively simple sensorimotor skills, and it's not clear that they'll scale up to explain the kind of stimulus-independent cognitive capacities, such as reasoning and planning, for which representational explanations have seemed most compelling. Moreover, proponents of these approaches often target their anti-representationalist arguments at the kind of syntactically structured symbolic representations invoked by early 'classicist' cognitive theories, yet neglect to identify fully *general* conditions for something's being a representation, so it's not clear whether their objections apply to representations as such, or merely representations of a certain kind.[1] Together these considerations motivate the following kind of worry: even if (say) a dynamical explanation of a stimulus-independent cognitive capacity were forthcoming, why not think that it simply provides an interesting new way of thinking about representation?

In his 2007 book, *Representation Reconsidered*, William Ramsey develops an anti-representationalist argument that seems to sidestep this sort of concern. He argues that the central threat to representationalism comes not from some imagined future non-representational cognitive explanation, but from existing theories of stimulus-independent cognitive capacities. He carefully examines the explanatory roles that

[1] For a canonical statement of the first sort of worry, see Clark and Toribio (1994). For a canonical statement of the second sort of worry, see Bechtel (1998)

representations have been taken to play within these extant theories and argues that some of these roles simply aren't representational in nature. In particular, he argues that theories within connectionism and cognitive neuroscience typically work by positing states that *detect* or reliably respond to some distal entity. These states, which Ramsey calls 'receptor representations', are characterized as representations by both connectionists and their most strident critics, but according to Ramsey there's nothing distinctively representational about the explanatory role that they play; all sorts of mechanisms function in essentially the same way without us having any inclination to call them representations. The only reason we continue to call receptors 'representations', claims Ramsey, is because of conceptual inertia left over from the cognitive revolution; just as cosmologists still *talked* about Ptolemaic celestial spheres even after the Copernican revolution rendered such spheres redundant, connectionists still *talk* about representations despite the fact that representations play no explanatory role within their theories.

Ramsey allows that *some* cognitive theories posit states that play a genuinely representational role. Specifically, he holds that 'classical' theories typically work by positing internal *models* that instantiate the same abstract structure as some external system, which are manipulated within the cognitive architecture so as to enable the agent to successfully reason about or interact with the external system. According to Ramsey, these models—or 'structural representations'—thereby function as genuine *stand-ins* for the external system. However, Ramsey holds that classical theories are on the decline in contemporary cognitive science, while broadly connectionist approaches are on the ascendancy.[2] Thus, he argues, the cognitive revolution is returning full-circle to a kind of neo-behaviorism: beneath the guise of representation-talk, contemporary cognitive science provides explanations in which "cognitive representations play no real explanatory role" (Ramsey 2007, p. 226). Further, while Ramsey makes no claims about whether connectionist explanations will ultimately prove to be *successful*, he argues that if they do, our commonsense conception of ourselves as intentional agents with internal mental states that are *about* the world, will have to be radically revised.

I don't think the distinction between structural representations and receptors lines up with the distinction between classicism and connectionism in the way Ramsey thinks it does. The idea that a trained neural network embodies the same abstract structure as its task domain is implicitly or explicitly reflected in a great deal of connectionist research, and, I think, has some claim to being the 'official' meta-theoretical view about how connectionist systems work. However, I will not press this point here, since several other authors have ably defended this line of response to Ramsey,[3] and although the response is sufficient to forestall Ramsey's anti-representationalist and eliminativist

[2] The term 'connectionism' is often used narrowly, to refer to a specific research program that emerged in the 1980s, which used highly idealized neural network models—typically, feedforward multilayer perceptrons trained by backpropagation—to simulate various psychological capacities. In this paper I'll use 'connectionism' more broadly, to encompass any psychological theory that appeals to signal processing within networks of nodes whose connections are shaped by experience-dependent plasticity mechanisms. Connectionism in this sense includes the PDP models of the '80s, as well as more biologically realistic models of specific neural circuits such as those found in contemporary cognitive and computational neuroscience.

[3] See Garzón and Rodriguez (2009), Grush (2008), Shagrir (2012), Sprevak (2011).

conclusions, I think there are deeper reasons why Ramsey's arguments for those conclusions fail, which have so far gone unnoticed, and which reveal a largely unexplored region of logical space in discussions about the explanatory role of representations.

In this paper I largely prescind from debates about whether or not connectionism posits structural representations, and focus on the cogency of the distinction between structural and receptor representations.[4] I begin, in Sect. 2, by discussing what Ramsey thinks it takes for a given type of theoretical entity to qualify as a cognitive representation. Then, in Sects. 2.1 and 2.2 turn to consider two purportedly different kinds of theoretical entity—structural representations and receptors, respectively—to examine whether they meet Ramsey's criteria for representationhood. In Sect. 3.1 I turn from exposition to criticism, and argue that when the two aforementioned concepts of representation are properly explicated, in ways that Ramsey is sympathetic to, there turns out to be no distinction between them; anything that satisfies the conditions of being a structural representation satisfies the conditions of being a receptor, and vice-versa. There only *appears* to be a distinction between structural and receptor representations because the former are tacitly conflated with introspectively accessible 'mental models' that plausibly mediate stimulus-independent cognitive capacities such as memory

[4] I should note that Ramsey discusses two other conceptions of cognitive representation that I will not discuss in this paper: 'input–output' (I–O) representations and 'tacit' representations. As with structural and receptor representations respectively, Ramsey holds that I–O representations are proprietary to classical explanations and play a genuinely representational explanatory role, whereas tacit 'representations' are proprietary to connectionist explanations and are not really representations at all. I do not discuss I–O or tacit representations at length in this paper because Ramsey himself places far more emphasis on the contrast between structural and receptor representations, and because the notion of structural representation is arguably the most important and widely discussed conception of representation in cognitive science. Moreover, I think that Ramsey's arguments about the representational status of I–O and tacit representations are far less convincing than his arguments about structural and receptor representations, though space limitations prevent me from giving anything more than a rough sketch of my reasons. First consider I–O representations. Ramsey holds that explanations in classical cognitive science proceed by first characterizing the cognitive capacity to be explained in terms of a mapping from inputs to outputs, which are characterized in terms of some external problem domain, and then decomposing the cognitive capacity into simpler sub-capacities, which are explained by appealing to computational sub-processes that implement 'internal' input–output mappings defined over the *same domain* as the overall capacity to be explained. Ramsey holds that these sub-processes therefore manipulate *epresentations* of entities within that domain. However, Ramsey's characterization of this explanatory strategy, which he identifies with the *homuncular functionalism* of Dennett (1981), strikes me as mistaken. The *whole point* of homuncular functionalism is that sub-processes do *not* manipulate representations of entities that are in the domain of the cognitive capacity to be explained—that's how decomposition is supposed to expunge the homunculus. Now consider tacit representations. Ramsey points out that connectionist explanations often invoke states that are somehow implicitly embodied throughout the functional architecture of a network, and holds that these 'tacit' states are characterized as representations merely because they dispose the network to settle into a certain pattern of activity. However, Ramsey argues, this kind of role isn't distinctively representational, for all sorts of physical states ground dispositions without us having any inclination to think of them as representations. The central reason this argument fails, I think, is that it rests upon a beguiling yet defective conception of explicitness—what Kirsh (1990, p. 350) has called the "bewitching image of a word printed on a page". Even a symbolic structure within a classical system, a paragon of computational explicitness, might be stored in essentially the same manner as the 'tacit' states of a connectionist network, in the sense that it might be arbitrarily distributed throughout memory, and only have a determinate identity by virtue of the way it is read by the processor—i.e. by virtue of the dispositions it grounds within the functional architecture of the system. Much more can and should be said about these issues, but unfortunately that will have to wait for another occasion.

and imagination. However, I argue in Sect. 3.2 that, contrary to widespread views in philosophy and psychology, there's nothing distinctively *mental* about structural representations, for they're to be found in all sorts of non-intentional systems such as plants. I conclude that although standard ways of explicating the nature of distinctively mental representations are unsuccessful, this doesn't license eliminativist conclusions; on the contrary, research into the mechanisms of 'mental models' is currently a thriving area of contemporary connectionism; philosophical views of mental representation just need to catch up.

2 The job description challenge

Suppose a renal physiologist told you that kidneys use representations to filter blood. *How*, you might reasonably ask. Upon being told about nephrons, antidiuretic hormones and such, you might reply: That's a perfectly good explanation of how the kidney functions to filter blood, but why should we consider that a *representational* function? That, surely, is the right kind of question to ask of representational claims about kidneys, and it's exactly the kind of question that Ramsey asks of representational claims in cognitive science. While cognitivist theories employ representation-talk, we shouldn't take that talk at face value if we're to understand the ontological commitments of such theories. Instead, we should ask whether the posits of those theories play a genuinely *representational* explanatory role. Of course, to answer such questions we need an account of what it is to play such a role. We need what Ramsey (2007) calls a *job description* for representations.

To ask what it takes for something to satisfy a representational job description is to ask the ontological question of what it takes for something to *be* a representation; but by framing this question in terms of job descriptions, Ramsey is indicating that he's adopting a particular methodological stance. He seeks not to develop a univocal analysis of representation that encompasses all and only the things we happen to call representations, for as he rightly points out, we apply 'representation' to such a congeries of things that such a project is surely forlorn. Instead, his project is located within the tradition of Quinean naturalistic ontology; he proposes that we look to specific cognitive theories and examine the explanatory roles that putative representations play within those theories. This approach leaves it open that those roles might fall under quite different types—that there might be quite *different* job descriptions for representations. It also leaves it open that a given explanatory role isn't genuinely representational after all. How might we decide whether a given role is representational? Ramsey holds that we must be guided at least to some extent by our intuitive, pre-theoretical notions of representation. While he doesn't endorse a fully descriptivist account of theoretical reference, he does hold that any proposed demarcation between those explanatory roles that are distinctively representational and those that are not must be continuous with our ordinary ways of thinking about representation, for otherwise the states that purportedly play a representational role would be representations in name only, like the alleged 'representations' in kidneys.

While Ramsey thinks that our commonsense notion of representation is a cluster concept, and hence isn't amenable to analysis, he thinks that three aspects of the notion are central, and impose (perhaps defeasible) criteria for something's satisfying a representational job description. First, and most importantly, representations *represent something*. We ordinarily think of both our representational mental states, such as beliefs and desires, and external representational artifacts, such as maps and words, as being *about* some object, property, or state of affairs. In the Brentanian jargon, we think of representations as being directed at an *intentional content*. Second, while this phenomenon of aboutness or intentional directedness seems to involve a kind of *relation*, if so, it must be a very *special* kind of relation, in which the distal relatum—the intentional object of the representation—needn't actually exist. To put it less tendentiously, it seems central to our conception of representations that they can *misrepresent*. Finally, we ordinarily think of the content of a representation as being somehow relevant to the causal role that the representation plays. If I represent someone as a friend, I might issue an affable 'Hello', but if I represent the person as my nemesis, I might engage in a very different, more sinister suite of behaviors.

Though Ramsey argues that we must *begin* with commonsense when identifying a job description for representations in cognitive science, he argues that we cannot *end* there, for although we have a venerable story about how external representational artifacts might misrepresent or play a role that's relevant to their content—roughly, they're interpreted by *us*—no such story will do for the sub-personal representations posited by cognitive scientists. The problem of course isn't just that there are no little homunculi interpreting our sub-personal representations, it's that positing such homunculi would lead to an explanatory regress. So showing that the sub-personal states posited by a cognitive theory satisfy a genuinely representational job description isn't just a matter of showing that they exhibit the various puzzling intentional properties of representations ascribed by common sense; it's also a matter of showing that the theory is able to explicate, in broadly mechanistic or 'naturalistic' terms, *how* those states exhibit those properties. This is what Ramsey calls the *job description challenge*.

Over the next two sections I'll look at two broad families of cognitive theories, and evaluate how they fare against the job description challenge. But before moving on, I'll say a few words to clarify the nature of that challenge. Through much of the late twentieth century, philosophers of mind sought to 'naturalize semantics' by providing informative, non-circular, and broadly 'naturalistic' conditions for an intentional state to have a particular semantic content—where having a content C is here understood as standing in a representation relation with C.[5] This project looms large over the philosophical landscape, so is apt to be mistaken for the job description challenge, however the two projects are quite different: one is about the metaphysics of the representation relation, whereas the other is about the ontology of representational vehicles. These projects are related, for as we've seen, one intuitive constraint on something's being a representation is that it has intentional content, but nevertheless they're distinct, and failing to recognize them as such can only lead to confusion.

As Ramsey and others have pointed out, this confusion is apparent in one of the most influential attempts to naturalize semantics: Millikan's (e.g. 1984) *teleoseman-*

[5] The main players in this project include Dretske (1988), Fodor (1990), and Millikan (1984).

tics. Millikan's view is complex and nuanced, but the central idea is that the content of a representational state is whatever has the biological function of eliciting that state. Ramsey points out that Millikan often seems to regard her view as an account of representation *as well as* an account of content, by counting any state that has the function of being elicited by something as a representation of that thing. For example, she holds that because adrenaline flow has the function of readying the body for situations that require strenuous activity, it *represents* such situations (Millikan 1984, p. 116). Yet it's not clear why readying the body for action is a distinctively *representational* function, any more than filtering the blood is. By thinking that her theory of content can do double-duty as theory of representation, Millikan seems to cast the net of representationhood too wide, over areas on which it has no explanatory purchase.[6] Indeed, Ramsey points out that this indifference to the question of what representations *are* is endemic throughout the literature on naturalized semantics, where theorists typically enquire into the conditions for something's having content, without asking what the nature of that something might *be*. Whether or not this gets the proper order of explanation backwards, or is simply a convenient way of carving up the problem space, the question of representationhood is clearly *distinct* from the question of content, and it's the former that we're interested in here.

2.1 Structural representations

Now we know the ground rules, we can judge how various competitors fare against the job description challenge. Let's begin by considering theories in the 'classical' tradition of cognitive science. Ramsey holds that when we examine classicist theories we find a recurring explanatory pattern: such theories typically attempt to explain cognitive capacities within a given task domain by positing neurally-encoded symbolic systems that instantiate the same abstract structure as some distal system within that domain. A given capacity is explained by showing how the internal symbolic system is computationally manipulated by the cognitive architecture so as enable the agent to successfully interact, in thought or action, with the external system. To borrow some helpful terminology from Swoyer (1991), the intuitive idea underlying this explanatory strategy is that, just as *we* use maps and scale models as 'surrogates' for reasoning about the real-world systems that those artifacts are structurally isomorphic with, the *brain* uses sub-personal cognitive models, encoded in symbolic systems, as surrogates for reasoning about the behaviorally-relevant distal systems that they're isomorphic with.[7] Following Ramsey, let's call the mechanisms that are hypothesized to play this kind of explanatory role *structural representations*.

[6] Others have made essentially the same point about the profligacy of Millikan's view. For example, Allen and Hauser (1993) complain that Millikan's view entails that "some interactions between trees can have content attributed to them" (p. 88), and Sterelny (1995) expresses concern that on Millikan's view, "it will turn out that saliva represents food" (p. 256).

[7] Note that I am using 'isomorphism' loosely here, to refer to the kind of resemblance relations that structural representations purportedly participate in, since that term is so familiar in this context. However, I will go on to argue that the resemblance relations at issue here are probably best understood as *homomorphisms* rather than isomorphisms.

The general idea that the mind reflects the abstract structure of external reality is prefigured in Aristotle's image of a signet ring impressing its form on a blob of wax. However, it arguably finds its most influential contemporary expression in Craik's (1943) seminal book, *The Nature of Explanation*. Craik suggests that,

> If [an] organism carries a 'small-scale model' of external reality and of its possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.
> (Craik 1943, p. 51)

Craik's suggestion is compelling, and seems to pick out an explanatory role which, if filled, would help explain a range of interesting cognitive phenomena. However, at the time he was writing it was difficult for psychologists to see how this role *could* be filled. The situation changed with the advent of cognitive science and its attendant notion of computation, for, as Fodor (e.g. 1985) is fond of pointing out, the notion of computation provided a way of understanding how an internal symbolic system might be manipulated, via purely mechanistic processes, so as to produce effects that are relevant to the semantic properties of the symbols. Add to this the idea that the internal symbolic system embodies the same abstract structure as some external system, and we begin to see how Craik's suggestion might be incorporated into a broadly mechanistic explanation of how the mind works.

Once cognitive scientists glimpsed the mechanistic bona fides of mental models, they invoked them freely to explain a wide range of cognitive phenomena. Perhaps the most prominent proponent of mental models was Johnson-Laird (1983), who drew direct inspiration from Craik and appealed to mental models to explain such sophisticated cognitive capacities as problem solving and deductive reasoning. However, the influence of the general idea that the mind manipulates internal structural isomorphs of external systems went far deeper, and some theorists suggested that all of the various representations posited by cognitive scientists ought to be unified under the aegis of the notion of structural representation; for example, Palmer (1978) *defined* cognitive representation in terms of a "correspondence (mapping) from objects in the represented world to objects in the representing world such that at least some relations in the represented world are structurally preserved in the representing world" (pp. 266–267). Similarly, many philosophers of cognitive science emphasized the importance of structural representations to explain cognition, especially the kind of 'sophisticated', stimulus-independent capacities discussed by Craik and Johnson-Laird; for example, Cummins (1994) claims that "what makes sophisticated cognition possible is the fact that the mind can operate on something that has the same structure as the domain it is said to cognize" (pp. 297–298).

Ramsey holds that mental models or structural representations are ubiquitous throughout *classical* cognitive science. He claims that they play an essential role within such quintessentially classicist theories as "Newell's production-based SOAR architecture (1990), Winograd's SHRDLU model (1972), Anderson's various ACT theories (1983), Collins and Quillian's semantic networks (1972), [and] Gallistel's computational accounts of insect cognition (1998)" (Ramsey 2007, p. 79; citations provided therein). However, Ramsey claims that structural representations are rarely invoked by *connectionist* theories.[8] As I mentioned in the Introduction, I think this claim is false. States that serve as models or internal isomorphs play just as important an explanatory role in connectionism as they do in classicism. One way to argue for this point is to show that many of the kinds of connectionist systems that Ramsey explicitly considers, such as the multilayer feedforward perceptrons discussed in the PDP literature of the '80s, in fact employ structural representations, despite Ramsey's claims to the contrary. Garzón and Rodriguez (2009) develop this line of objection to Ramsey by pointing out that techniques like cluster analysis show that the metric relations between hidden-layer activation patterns in a trained multilayer perceptron reflect relations between entities in the problem domain that the perceptron was trained to classify. Indeed, this is the standard interpretation of how perceptron classifiers work, and is widely endorsed by the most prominent proponents of connectionism. As the Churchlands Churchland and Churchland (2002) put it, "The various distance relationships between the learned clusters. . .within the activation space of a given population of neurons, are collectively and literally isomorphic with the similarity relationships that objectively exist between the various categories in the external world" (p. 907).

Ramsey doesn't directly address the point that even canonical connectionist systems are standardly characterized as vehicles of structural representation, but some of his related comments suggest the following reply: any structural relations between activation patterns purportedly revealed by cluster analysis are merely dispositionally latent within the network, and therefore don't reflect structural relations in the task domain.[9] However, some have made similar claims about the structural relations purportedly encoded by *classical* systems. For example, quoting Cummins (1989), O'Brien and Opie (2001) suggest that conceptions of representation based on structural resemblance have no place in classicism since "nothing is more obvious than that [symbolic] data structures don't resemble what they represent" (Cummins, 1989, pp. 30–31). The problem with the general claim here, whether it's made on behalf of connectionists or classicists, is that it rests on a mistaken 'pictorialist' conception of structural resemblance. When considering whether a dynamic physical mechanism serves as a model, we must look to the functional, dispositional properties of the mechanism, not to its static, categorical properties. As Craik (1943) insightfully put it, "a model need not resemble the real object pictorially; Kelvin's tide predictor, which consists of a number of pulleys on levers, does not resemble a ride in appearance, but it works in the same way in certain essential respects" (p. 51). What matters is not how the mechanism appears to the human eye, but how the mechanism *works*, and cluster analysis seems to show that classifier perceptrons work by instantiating the categorical structure of the domain that they classify.

[8] Where I'm here using 'connectionism' in the broad sense I outlined in note 2.

[9] This reply, and my response to it, echo many of the points about so-called 'tacit' representations that I address in note 4.

A second way to show that structural representations play an important explanatory role in connectionism is to consider some less familiar connectionist models, and argue that they too invoke structural representations. This strategy is pursued by Shagrir (2012), who discusses recent models of oculomotor control in computational neuroscience, which posit stable states of reverberating activity in a recurrent neural network that can occupy various points along a line attractor. The current location of the attractor state serves to keep track of the current position of the eyes. According to Shagrir, the network functions as a structural representation since "the state-space of the network mirrors the space of eye positions" (*ibid*., pp. 13–14). When evaluating the significance of Shagrir's example, it's important to keep in mind that Ramsey allows that *some* connectionist theories invoke structural representations; he cites Grush's Grush (2004) 'emulation theory' as an example. Ramsey's claim is that such theories are recherché exceptions to a general trend of explaining cognition without appealing to structural representations. Shagrir is perhaps insufficiently attentive to this point, so it's important to note that the oculomotor model he discusses is an instance of a general class of attractor network models that is widely invoked throughout contemporary computational neuroscience to explain a diverse range of psychological phenomena,[10] and the points he makes about the oculomotor model could be generalized to other models in the same class. Similarly, as Grush would be the first to admit, his 'emulation theory' did not spring fully formed from the head of Zeus, but rather offers a synthesis of existing work on the neuroscience of motor control; 'emulator' is effectively Grush's term for a kind of theoretical posit—a forward model—that is ubiquitous in mainstream motor neuroscience.[11] The theories mentioned here are far from recherché; they're about as central to contemporary neuroscience as can be.

The general line of response to Ramsey that I've been sketching here, according to which structural representations are not propriety to classicist theories, has been developed by several other authors in addition to those I've already mentioned,[12] so I will not pursue it further here. Instead, I will pursue an alternative line of response, which to my knowledge has not been explored in the existing literature. My central goal in this paper is not to adjudicate whether classicism or connectionism has custody over structural representations, but to diagnose whether the notion of structural representation is legitimate; I will argue that there is simply no distinction between structural representations and the purportedly non-representational 'receptors' that Ramsey contrasts them with.[13]

To develop this argument, I should first consider whether structural representations satisfy a representational job description. Ramsey holds that they clearly *do*, since their explanatory function is to *model* external systems. Like the scale models used by, say,

[10] See Eliasmith (2005) for a review.

[11] See Miall and Wolpert (1996) for a review.

[12] See also Grush (2008), Sprevak (2011).

[13] Note that the boundary between these two kinds of replies to Ramsey is fuzzy. Some of those who have argued that connectionists invoke structural representations can be understood to be arguing, implicitly, that states that intuitively seem to be receptors in fact also count as structural representations. However, to my knowledge nobody has explicitly developed an argument to the effect that all and only structural representations are receptors.

aeronautical engineers, they are intuitively *about* the systems they are structurally isomorphic with. Moreover, the fact that they are isomorphic with a given system is relevant to their causal role, since it's in virtue of that structural resemblance that they can be manipulated so as to guide an agent's successful interactions with that system. However, unlike the scale models used by engineers, they needn't be interpreted by an intelligent agent to play the content-relevant, behavior-guiding causal roles that they do; insofar as they can be manipulated by purely formal, computational operations, there's no metaphysical mystery about how they could play those roles in the absence of a homunculus.

To evaluate Ramsey's sanguine attitude about the job prospects for structural representations, I think we need a much clearer view of what exactly structural representations *are*. This is especially urgent given that, as Ramsey rightly points out, structural representations seem to be a species of what Peirce called *icons*—entities that represent by virtue of *resembling* what they are about—and there are of course venerable objections to resemblance-based theories of representation. One of the most classic such objections is directed specifically at the view that mental representations resemble their intentional contents. Surely, the objection runs, my mental image of, say, an orange isn't itself colored orange.

This objection seems decisive against a resemblance-based theory of representation that explicates resemblance in terms of the co-instantiation of 'first-order' monadic properties. However, structural views of representation sidestep this objection by cashing resemblance out in terms of the co-instantiation of 'second-order', *relational* properties (O'Brien and Opie 2001; Shepard and Chipman 1970). A canonical example of an iconic representation that represents in this sense is a cartographic *map*: the map instantiates the same relational (geometric or topological) structure as the terrain it represents. However, as I suggested earlier, the functional architecture of a mechanism might also embody the same relational structure as an external system—and indeed, it's arguably this kind of *dynamic* rather than *pictorial* structural resemblance that's important when we're discussing models that can function without being interpreted by an intelligent homunculus.

Ramsey, like many other proponents of structural representation, attempts to make the notion of shared relational properties more precise by appealing to the mathematical notion of *isomorphism*. However, a notion of resemblance cashed in terms of isomorphism is still a notion of resemblance, and Goodman (1968) famously developed what many take to be decisive objections to *any* kind of resemblance-based theory of representation. His central argument is that the logical properties of resemblance relations seem utterly different from those of representation relations: resemblance is reflexive, symmetric, and transitive, whereas representation is none of these.

To avoid the brunt of this objection, some theorists have attempted to explicate structural representation in terms of the notion of *homomorphism* rather than isomorphism.[14] Homomorphisms are a more permissive kind of structure-preserving mapping, which can obtain between systems with different cardinalities, and which

[14] See, e.g., Bartels (2006), who develops this idea in the context of debates about scientific representation, a context in which many of the present issues about structural representation are recapitulated.

needn't be either symmetric or transitive.[15] Cashing out a notion of structural representation in terms of homomorphisms has also seemed attractive because it allows that representing systems might be less than perfect simulacra of represented systems.

Appealing to homomorphisms appears to resolve *some* of the problems with an unvarnished resemblance-based theory of representation. However, it still leaves us with the problem that resemblance is reflexive. Intuitively, representations don't represent *themselves*. There's an arguably deeper problem here, too: abstract structural relations are *ubiquitous*. Slough off enough detail, and virtually anything might be homomorphic with virtually anything else. A map of the NYC subway system might reflect the topology of connections between actual subway stops in Manhattan, but it might also reflect the topology of connections between a population of neurons in your brain. Yet the map surely doesn't *represent* your neurons. In the case of artifacts like subway maps, the relevant homomorphisms are plausibly constrained by interpretative intentions of agents; but how do we constrain the homomorphisms that cognitive representations participate in without appealing to homunculi?

Ramsey's response to this problem is puzzling. He holds that the morphisms that constitute structural representation relations are constrained by the *methodology* of cognitive explanation: "the explanandum itself…determines what it is that is being modeled" (Ramsey, 2007, p. 94). So Ramsey seems to think that if, for example, a cognitive theory posits an internal map to explain a rat's ability to navigate home, the map represents spatial relations in the rat's environment—as opposed to, say, some configuration of stars in Andromeda—because the theory is about the rat's ability to navigate its environment (*ibid.*, p. 96). This seems to entail that the content of structural representations is radically observer-dependent; that, for example, rats had to wait until they were studied by human ethologists until their cognitive maps acquired determinate content. One might think this constitutes a *reductio* of Ramsey's view.[16]

Ramsey does hint at an alternative view, though he doesn't distinguish it from the one just criticized, namely that the relevant morphisms are constrained by *vcausal relations* between representing and represented structures. On this view, the rat's cognitive map represents the geometry of its environment at least in part because the map is causally responsive to specific geometric properties of the environment. A second

[15] An isomorphism is a bijective (i.e. one-one and onto) function from one set-theoretic structure to another, which preserves the relations defined over the elements of each structure. More precisely, an isomorphism between structures A and B is a bijective mapping $\phi : A \rightarrow B$ from the objects in $A = \{a_1, \ldots, a_n\}$ to the objects in $B = \{b_1, \ldots, b_n\}$, such that for any relation $R \in A$, if R obtains for a subset of the objects in A, $A' = \{a_i, \ldots, a_j\}$, there is a relation $S \in B$, that obtains for a subset of the objects in B, $B' = \{\phi(a_i), \ldots, \phi(a_j)\}$. A homomorphism, like an isomorphism, is a structure-preserving mapping from one set-theoretic structure to another, but unlike isomorphisms, homomorphisms needn't be bijective. Thus, a homomorphic mapping can be *many-one*, and needn't map *onto* all the of the elements in the represented structure.

[16] Ramsey (2007, pp. 98–99) does respond to a *kind* of observer-dependency worry in this context, but not the specific worry that I'm raising here. He addresses the concern that the structural representations posited by classical cognitive scientists are merely useful fictions. Like Ramsey, I don't think we should lose any sleep over *that* concern. The worry that I'm raising here is different: it's that *Ramsey's explication* of the kind of structural representations posited by classical cognitive scientists entails that the content of a structural representation is radically observer-dependent. One might be troubled by that worry without having any specific views about the scientific realism debate.

way in which causation might help constrain the relevant morphisms is via the *causal role* that the representing system plays within an agent's cognitive architecture; the idea here is that the representing system is *used* to guide the agent's interactions with the represented system. This sort of use-condition is widely invoked in discussions of cognitive representation,[17] and is clearly related to the idea that a representation ought to play a causal role that is relevant to its content; but it might also help with the indeterminacy problem presently under consideration. To paraphrase Dretske (1988), it is only by *using* a representation in the production of movements whose successful outcome depends on *what is being represented* can indeterminacy about the target of the representation be overcome (p. 70). So appealing to external causal relations, or internal causal roles, seems to make headway on the problem of the ubiquity of structural resemblance. It also seems to resolve the problem of the reflexivity of resemblance. For while it's plausible that a rat's cognitive map is causally responsive to the geometric properties of the rat's environment, the reverse is surely not.

Over the past 30 years or so, the psychologist C. Randy Gallistel has championed an account of cognitive representation that incorporates the various advances over a simplistic resemblance theory that I've just surveyed.[18] Gallistel's account thus promises a precise articulation of the kind of structural view of representation that Ramsey gestures at. According to Gallistel, a system A counts as a representation of a system B just in case it satisfies three conditions: first, A is homomorphic with B; second, this homomorphism is established and sustained by causal relations between the two systems, such that variations in A are causally influenced by variations in B; and third, A causally interfaces with motor control systems such that it can guide the agent's behavior with respect to B in ways that reflect the relevance of B for the agent. Thus, for Gallistel, representation relations are *functioning homomorphisms*, that is: abstract structural similarities, sustained by causal relations, which serve to inform an agent's behavior.

Gallistel's account is clearly an instance of a structural view of representation in Ramsey's sense, and Ramsey himself characterizes Gallistel's account as an exemplar of a structural view (recall the quote on page 9). It's true that there are some differences of emphasis between the two authors' accounts—for example, Ramsey tends to focus on the role of structural representations in surrogative reasoning, whereas Gallistel emphasizes that they are used to "control and direct appropriate behavior" (Gallistel and King 2010, p. 55)—however these differences are superficial. For example, Gallistel sometimes expresses his view in terms of surrogative reasoning. In a (1990) paper, he writes that "In order for a representation to exist, the neural or mental representatives of environmental variables must enter into combinatorial neural or mental processes that generate *valid inferences* about the represented variables" (p. 4, my emphasis). Conversely, Ramsey's conception of surrogative reasoning is very liberal, and seems more or less coextensive with the kinds of behavioral control processes emphasized by Gallistel. As I'll discuss in more detail later, one of Ramsey's central examples of a structural representation is the mechanism depicted in Fig. 1 below:

[17] See, for example, Godfrey-Smith (2006), Grush (2004), and Millikan (1984).

[18] The most detailed presentation of his theory of representation appears in his most recent book, Gallistel and King (2010).

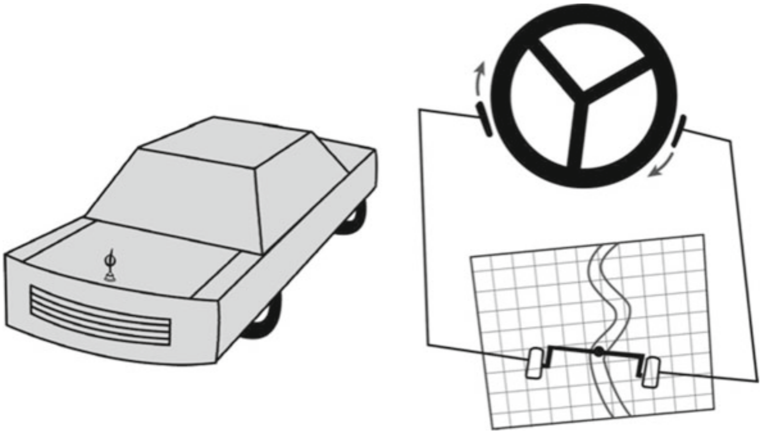


Fig. 1 Ramsey's example of a structural representation: An S-shaped groove inside a toy car interfaces with the car's driveshaft and steering wheel in such a way that when the car moves forward, a rudder traces the course of the groove, thereby turning the wheel and enabling the car to navigate an S-shaped road. The S-shaped groove embodies the same abstract structure as the road, and serves to guide the car along the road, hence satisfies Ramsey's criteria for structural representation (from Ramsey 2007, p. 199).

an S-shaped groove inside a toy car, which interfaces with the car's driveshaft and steering wheel in such a way that when the car moves forward, a rudder traces the course of the groove, thereby steering the car and enabling it to navigate an S-shaped road. Ramsey holds that the car is able to successfully 'reason' about the road using the groove as a surrogate, by virtue of the structural resemblance between the groove and the road. Now, one might quibble about whether a process like this is properly called 'reasoning', but what's important to note here is simply that both Ramsey and Gallistel, a crucial condition for something's being a structural representation is that it's *used* in the right way.

Despite these basic similarities, we've seen in this section that there are some lacunae in Ramsey's attempt to articulate a notion of structural representation that avoids the traditional problems with resemblance-based theories of representation. Gallistel's notion of a functioning homomorphism seems to fill these lacunae, and hence appears to provide the strongest version of a structural conception of representation on offer.[19] Thus in what follows I'll take 'structural representation' to be coextensive with whatever satisfies Gallistel's conditions for participating in a functioning homomorphism.

2.2 Receptor representations

Let's now turn to an apparently very different kind of state that has been invoked to explain cognition, and ask whether it satisfies a job description for representations.

[19] Note that I don't take myself to have provided a strong defense of a Gallistelian view of structural representation against the traditional objections to resemblance-based theories. For such a defense, see Isaac (2012). My claim here is simply the conditional one that if any account of structural representation is viable, Gallistel's is the most plausible candidate.

Ramsey argues that when we look at the family of theories under the broad banner of 'connectionism', we find that such theories routinely appeal to states that are said to represent events of a certain type by virtue of being reliably elicited by, or correlated with, such events. We see this most clearly in neurophysiology, when individual neurons are characterized as 'detectors' of the class of entities that reliably elicit maximal activation within that neuron. An example that's perhaps most familiar to philosophers is the 'bug detector' in the retina of the frog described by Lettvin et al. (1959), but similar examples abound within neurophysiology; one finds edge detectors (Hubel and Wiesel 1962), face detectors (Desimone 1991), even particular spatial location detectors (O'Keefe and Dostrovsky 1971). Ramsey argues that we find essentially the same idea in the PDP literature of the '80s, where stable patterns of activity in trained multilayer perceptrons are said to *represent* the inputs that they were elicited by. Whether the focus is on activity in single neurons, or in neural networks, the explanatory pattern seems to be the same: internal states are said to represent distal events by reliably 'detecting' them.

Ramsey calls the states that play this kind of explanatory role 'receptor representations'. However, he argues that they are not representations properly so-called, for the role they play is not a genuinely representational one. Simply functioning as a reliable receptor cannot be *sufficient* for something to count as a representation, claims Ramsey, since all sorts of mechanisms function as receptors without us having any inclination to think of them as representing anything. We already saw one example when discussing the profligacy of Millikan's view: adrenaline flow is reliably elicited by stressful events, but there's arguably no strong sense in which it *represents* such events. Further examples abound: antibodies in the immune system, bimetallic strips in thermostats, infrared sensors in automatic faucets, and so-on. All of these things work by entering into particular states in response to certain distal conditions, but none of them function *as* representations by doing so. They simply function as triggers, or relays, or causal mediators.

The last example here provides a stark illustration of Ramsey's point. While automatic faucets might be preferred over manually-controlled ones for various reasons, the role that infrared sensors play in automatic faucets isn't any different from the role that handles play in manual ones; both mechanisms just serve to turn the water on. We might adopt the intentional stance with respect to automatic faucets and interpret them as 'recognizing' the hands that trigger them, but this gloss is superfluous when it comes to understanding how automatic faucets actually work. Ramsey holds that the same is true of the purportedly representational states posited by connectionists. Unlike structural representations, we can understand the explanatory role that these so-called receptor representations play without understanding them as inner stand-ins for external entities. To put the point another way, any content we might ascribe to receptors is irrelevant to explaining the role that the receptor plays within the system of which it is a part.

At this point one might worry that the preliminary characterization of receptors that we've been working with so far is a bit too thin to capture the notion of representation that connectionists have in mind when they describe feature detectors or states of networks as representations. For it seems that virtually *any* state that's causally dependent on another counts as a receptor representation according to this initial characterization.

Yet surely the notion of representation that connectionists employ when they claim, for example, that place cells represent a particular spatial location is somewhat more constrained than the idea that, say, a shattered window pane 'represents' the rock that broke it. The question that presents itself, then, is whether we can discipline the notion of receptor representation in such a way that it doesn't encompass just *any* causal relation, yet retains the character of 'detection', while also satisfying the job description challenge.

One might have thought that the literature on indicator semantics could provide a promising source of constraints, since the central project in that literature is to explain intentional content in terms of reliable causal dependencies, while avoiding the obvious problem that not all causal dependencies involve intentional content. However, as I mentioned earlier, proponents of indicator semantics tend to ignore questions about what representations are and how they function. One notable exception is Dretske (1988), whose account of content is largely motivated by concerns about the explanatory role of representations. So, following Ramsey, let's consider whether Dretske's account might help to discipline the notion of receptor representation so that it passes the job description challenge.

Dretske's account of intentional content is driven by two central ideas. The first, common to all versions of indicator semantics, is that the relation between intentional states and their contents is fundamentally a kind of *causal* relation. In addressing the aforementioned problem that not just any causal relation involves intentional content, Dretske also attempts to do justice to a second idea about the nature of intentional content, which we encountered earlier when discussing the job description challenge: that intentional states play causal roles that are somehow relevant to their content. As Dretske (1988) puts it, "The fact that [intentional states] have a content…must be relevant to the kind of effects they produce" (p. 80). Dretske brings these ideas together by holding that a given state R that is causally dependent on some type of external event C comes to have the intentional content that C when R is selected, *via* natural selection or individual learning, to play a specific causal role within the system of which it is a part, *by virtue of the fact* that it is causally dependent on C.[20] That is, R is about C not only when it responds to C, but when it has the *function* of so responding, by virtue of the fact that its responsiveness to C is exploited by the system so as to guide appropriate behavior with respect to C.

Dretske's account seems to provide a characterization of receptor representations that doesn't massively over-generalize in the way that the original characterization did. To begin to see why, consider the 'bug detectors' in frogs described by Lettvin et al. (1959). The researchers described these cells as *bug* detectors, rather than *small-dark-moving-dot* detectors, even though they showed that the cells are maximally responsive to small dark moving dots. This suggests that the researchers were employing a more robust notion of representation than one according to which a mechanism represents whatever it is causally responsive to. Given the presumably uncontroversial assumptions that the frog's bug detectors interface with mechanisms that mediate bug-directed

[20] The distinction between an indicator's being selected over the course of phylogeny, and its being selected over the course of ontogeny, plays an important theoretical role for Dretske. However the distinction is largely orthogonal to present debates, so, like Ramsey, I'll elide over it in what follows.

tongue strikes, and that they have been selected to do so *because* they reliably respond to bugs, Dretske's account of content—taken as an account of representation—seems to capture this more robust notion; it seems to provide a sense in which bug detectors genuinely detect bugs *as such*. Moreover, Dretske's account would presumably encompass many of the other kinds of states claimed by connectionists to be representations, yet would exclude shattered window-panes and the like. So Dretske's account seems to provide a robust explication of the notion of receptor representation.

But do receptors in Dretske's sense satisfy a representational job description? At first glance it would seem that they do. Indeed, Dretske expressly develops his account in order to capture the three aspects of our commonsense conception of representation discussed in Sect. 2. However, Ramsey argues that appearances here are illusory. He points out that the fact that a structure is responsive to certain conditions, and that it is enlisted to play a particular functional role within a system by virtue of that responsiveness, doesn't entail that the structure plays a distinctively *representational* role, for "a structure can be employed... *qua* reliable respondent without being employed... *qua* representation" (Ramsey, 2007, p. 138). For example, he points out that one might plant a tree at a particular location to provide shade at a particular time of the day, exploiting the fact that the shadow of the tree depends on the position of the sun. However, this dependency is used to provide shade, not to provide information about the position of the sun. Of course, causal dependencies might *sometimes* be exploited for representational purposes. You might use the tree as a sundial, taking the shadow it casts to inform you about the sun's position, and hence the time of day. Ramsey's point, however, is that not all cases of exploiting a causal dependency are cases of exploiting a causal dependency *for representational purposes*. So, he argues, merely satisfying Dretske's conditions isn't sufficient for something to count as a representation. Even if we make receptors more robust by appealing to Dretske's conditions, it seems that they still fail to meet the job description challenge.

3 Problems in the workplace

I find Ramsey's anti-representationalist argument unconvincing. To begin to see why, first notice that even if we endorse all the premises of the argument, the conclusion simply doesn't follow. Ramsey wants to argue that the 'receptor' mechanisms typically characterized as representations by connectionists—such as feature detectors, states of networks, and the like—aren't really representations at all. But notice that all that the argument at the end of the preceding section entitles us to conclude is that such mechanisms aren't representations *simply by virtue of being indicators that are exploited for behavioral control*. This doesn't entail that receptor mechanisms aren't *representations*; it just entails that insofar as they are representations, they must be so by virtue of satisfying certain additional conditions.[21].

[21] Sprevak (2011) makes essentially the same point when he writes that "what satisfies the receptor notion, by itself, may not fulfill the job description of a representation, but the wider explanatory role that it plays in explaining successful behaviour may justify its labelling as a representation."

To illustrate, consider a case that parallels Ramsey's example of using a tree for shade. The brains of fruit flies contain neurons that are responsive to temperature, and these so-called 'thermosensors' allow fruit flies to avoid extreme heat (Hamada et al. 2008). Like the shadow cast by Ramsey's tree, the causal dependency of thermosensor activity on temperature is exploited for the purpose of staying cool—but it doesn't follow from this that the dependency is not *also* exploited for informing the fruit fly about the location of heat sources. Similarly, I might read the weather report to decide whether to carry my umbrella, but it doesn't follow that the weather report isn't also informing me about the likelihood of rain. Now, I haven't yet provided a strong reason to think that receptors like thermosensors *are* representations; my point is simply that Ramsey's argument, taken on its own terms, doesn't provide a strong reason to think that receptors are *not* representations.

3.1 Structural receptors?

However, I think there are more fundamental reasons why Ramsey's argument fails. We can begin to uncover these reasons by first noting that the length of the tree's shadow in Ramsey's example is *homomorphic* to the angle of elevation of the sun above the horizon. That's why we could use the length of the shadow to *measure* the sun's elevation; according to standard ways of thinking about measurement, the possibility of measurement presupposes the existence of a homomorphism from the measured into the measuring structures (Krantz et al. 1971). Now, if we *were* to use the length of the shadow to measure the sun's elevation, would we be using the shadow as a structural representation or as a receptor? There's a sense in which the shadow's length would be serving as a kind of *model* of the sun's elevation for us, yet the shadow would also provide a way for us to *detect* the sun's elevation; indeed, it's precisely because the shadow 'detects', or covaries with, the elevation of the sun, that it could be used to measure the sun's elevation. Further, supposing that there *is* a robust sense in which the shadow in this case would be serving as a structural representation, why wouldn't Ramsey's objection to the representational bona fides of a Dretskean conception of receptors carry over to it? That is, why couldn't we object, by parity with Ramsey's objection to receptor representations, that if one were to use the tree's shadow to stay cool, one would be exploiting a causally-mediated homomorphism for a specific purpose—albeit a *non-representational* purpose—and hence that merely satisfying the conditions for being a functioning homomorphism isn't sufficient for something to play a distinctively representational role?

Ramsey might reply that this question betrays a misunderstanding of the kind of explanatory role that is constitutive of structural representationhood; structural representations are supposed to be homomorphisms between a representing system and a represented system that can be exploited to successfully guide an agent's inferences or behavior *with respect to the represented system*. Arguably, that's not happening in the case of using the tree's shadow for shade—you're not using the shadow to inform your behavior with respect to the sun's position, you're just using it to stay cool. So perhaps this case doesn't satisfy the conditions for structural representationhood after all. But why couldn't a proponent of receptor representations reply in a precisely parallel fashion? Indeed, in criticizing the Dretskean reconstruction of the receptor notion of representation, Ramsey seems to overlook a central tenet of Dretske's theory. According to Dretske, indicators of C become representations when they are enlisted to interface with certain motor commands *by virtue of the fact* that their being so enlisted leads to successful behavior with respect to C. As Dretske (1988) puts it, "Only by *using* an indicator in the production of movements whose successful outcome depends on *what is being indicated* can [indeterminacy about the target of a representation] be overcome" (p. 70).

If proponents of structural representation can appeal to this sort of use-condition to deflect the shadow-used-as-shade counterexample, why can't proponents of receptor representations? It certainly seems plausible that neuroscientists have this sort of condition in mind when, for example, they characterize neurons that respond to small, dark moving dots as *bug detectors*; plausibly, such neurons have been selected to control tongue-snapping *because* they enable successful bug-catching. Perhaps one might be dissatisfied with the present line of response to the shadow-used-as-shade counterexample; however, it's not clear why one would be any *more* satisfied with the response in the case of structural representations than in the case of receptors—or vice-versa. So it's beginning to appear that structural representations and receptors are *on all fours* when it comes to meeting the job description challenge.

Now that I've exposed some raw concerns about Ramsey's arguments, I'll refine them into a more substantive objection. This will involve two main steps. First, note that many of Ramsey's canonical examples of receptors involve mechanisms whose states vary in proportion to the magnitude of some distal quantity. For example, the conformation of the bimetallic strip in a thermostat varies in proportion to the ambient temperature. Many of the receptor-like mechanisms posited by neuroscientists are similar in this respect. For example, the activity of thermosensors in fruit flies also varies in proportion to the ambient temperature. Call such mechanisms *analog receptors*.[22] As I suggested above, it seems clear that analog receptors participate in homomorphisms, from the various possible magnitudes that the receptor is responsive to, to the various possible states of the receptor mechanism; that is, relations between different different states of the mechanism *model* relations between different magnitudes. I see no reason to deny this; indeed, it strikes me as an uncontroversial application of conventional ideas about the nature of measurement. It's difficult to see how one could deny that analog receptors participate in homomorphisms without also denying that, say, thermometers are measuring instruments. Moreover, the sense in which analog receptors participate in homomorphisms seems precisely the same sense in which structural representations do; indeed, the psychologists who have done most to explicate the theoretical underpinnings of the notion of structural representation emphasize

[22] As I'm using the term here, 'analog' is not synonymous with 'continuous'. Rather, I'm using 'analog' in roughly Lewis (1971) sense, according to which magnitudes in the representing system are directly proportional to magnitudes in the represented system. However, when I go on to distinguish between 'analog' and 'binary' receptors, the distinction I wish to mark is not Lewis' distinction between analog and digital representation. Binary receptors are arguably still cases of analog representation in Lewis' sense, it's just that they're degenerate cases; they can occupy only one of two states.

the essential similarity between that notion and the notion of representation found in measurement theory (Gallistel 1990; Palmer 1978).[23]

Now, it's true that not all of the mechanisms characterized as receptors by Ramsey obviously count as *analog* receptors. For example, as their name suggests, *feature detectors* have traditionally been characterized as having the function of detecting the presence of a given property, rather than of covarying with the magnitude of some quantity.[24] It's worth noting that this interpretation of feature detectors is in fact highly controversial. Problems with this interpretation were revealed by a neural network model that was trained to extract information about the shape of an object from patterns of shading (Lehky and Sejnowski 1988). The network contained units that came to have response properties that were remarkably similar to the 'edge' and 'line' detectors in the visual cortex, yet an analysis of the role that the units played within the network suggested that they served not to detect contours, but to extract curvature information. The authors drew the general lesson that to understand the function of a feature detector, we must look beyond its response profile to the entire role it plays within the network in which it is embedded. Consonant with this suggestion, an emerging consensus in computational neuroscience suggests that feature detectors serve to communicate the degree of error between feedforward sensory information and feedback predictions.[25] From this perspective, feature detectors *do* serve to covary with magnitude of some quantity, hence count as analog receptors, so all the points made in the previous paragraph apply to them.

However, even if we interpret feature detectors traditionally, as what we might call *binary receptors*, there's still a robust sense in which they participate in homomorphisms: relations between states of the mechanism reflect relations between states of the system that the mechanism is responsive to[26]. Such homomorphisms might be very *simple*, but they're homomorphisms all the same. Similarly, mechanisms like oil lights on car dashboards, smoke alarms, and pregnancy tests might participate in very simple homomorphisms with the systems they measure, but they nevertheless *do* participate in homomorphisms with those systems; that's how they're capable of measuring them. An oil light is no less a measuring instrument than a fuel gauge, simply because it can occupy fewer states; it's just a less *discriminating* measuring instrument.

[23] Gallistel (1990) writes that "representation should have the same meaning in psychology as it has in mathematics" (p. 1), and that "those familiar with the theory of measurement. . .will recognize the parallel between this use of representation and its use in measurement theory" (p. 2).

[24] The idea, of course, is not that a feature detector fires *when and only when* the feature that it is tuned to is in its receptive field; neurons are noisy critters, and are constantly firing even in the absence of external stimulation. The idea is that although a feature detector might fire across a range of frequencies, it's only firing above a certain threshold frequency that is functionally relevant to its role as a detector.

[25] A seminal application of this idea to 'edge' and 'line' detectors is Rao and Ballard (1999).

[26] This might be a bit quick. Some, such as van Fraassen (2008), argue that there's simply no sense to be made of the idea that a homomorphism might hold from one concrete, physical system to another, since the technical notion of a homomorphism is only well-defined in the domain of abstract, mathematical systems. This issue deserves further discussion, but it's orthogonal to my present concerns, since it doesn't provide any grist for the mill of someone who wants to claim that there's a substantive theoretical distinction between what I'm calling 'analog' and 'binary' receptors, or between receptors and structural representations. Insofar as the notion of homomorphism applies to *any* of these mechanisms, it applies to *all* of them.

The essential point underlying this discussion is that homomorphisms are *flexible*. Earlier we saw one dimension along which they're flexible: 'horizontally', they might apply between virtually any two systems. We saw that this dimension of flexibility poses a prima facie problem for theories that attempt to explain representation in terms of homomorphisms, but that the problem can arguably be assuaged by appealing to external causal relations and internal causal roles. However, the present discussion has revealed that even if we restrict our attention to functioning, causally-mediated homomorphisms, homomorphisms are still flexible along a second dimension: 'vertically', they might hold at many different levels of granularity. This is because homomorphisms exist wherever there is a relation of reliable causal covariation that might be exploited for measurement purposes, and such relations exist at many different levels of granularity—a mechanism that covaries with some system might be highly articulated and capable of embodying fine-grained distinctions, or very simple and only capable of making binary distinctions. While this flexibility needn't pose a *problem* for a structural theory of representation, it does begin to suggest that such a theory will be far more liberal than its proponents might have initially thought.

That was the first step in my objection to Ramsey's anti-representationalist argument. The second is to observe that, aside the emphasis on homomorphisms, Gallistel's 'structural' theory of representation is essentially just a notational variant of Dretske's 'receptor' theory. Both authors emphasize that cognitive representation is fundamentally a matter of exploiting a causal relation between representing and represented systems so as to guide successful interactions with the represented system. Recall from the discussion at the end of Sect. 2.1 that although both Ramsey and Gallistel sometimes characterize structural representations in terms of *surrogative reasoning*, neither of them thinks that such 'reasoning' need be especially sophisticated. For both authors, what matters is that the represented system is *used* in the right way; namely, to guide successful interactions with the represented system. But we've seen repeatedly now that Dretske emphasizes essentially the same point. If there's a substantive difference between Gallistel's view of representation and Dretske's, it's that the former requires that representing systems be *homomorphic* with represented systems. Yet we just saw that this isn't a substantive difference at all; insofar as changes in representing systems reliably covary with changes in represented systems, the homomorphism requirement will be satisfied. Homomorphisms are vertically flexible. It's true that the kinds of mechanisms we *intuitively* associate with a Dretske-style, receptor-based notion of representation are relatively simple analog or binary receptors. However, there's no reason such a notion shouldn't encompass richer mechanisms that are intuitively more model-like. And in any case, when we focus on the *theoretical content* of the receptor- and structural-based notions of representation, rather than the connotations of words like 'detector' and 'model', it seems that whatever satisfies the conditions of one notion will satisfy the conditions of the other.

Let's see how these points play out in the context of a specific example. The physiology and behavior of almost all animals on Earth is modulated by a circadian rhythm that reflects the period of the Earth's axial rotation. These rhythms are mediated by *circadian clocks*: biochemical mechanisms that produce an endogenous oscillation of roughly 24 h, which is entrained to the phase of the Earth's day-night cycle by *zeitgebers*, or daylight cues. Circadian clocks play a crucial role in regulating a range of metabolic processes and behaviors that depend for their success on being synchronized with the day-night cycle. Thus circadian clocks seem to satisfy Gallistel's criteria for representationhood: they enter into homomorphisms with external systems; these homomorphisms are established and sustained by causal relations; and they are exploited in the service of behavioral success. Indeed, Gallistel (1998) agrees: "The circadian clock. . .is perhaps the simplest of all the well-documented functioning [homomorphisms] between brain processes and the world" (p. 28).[27] However, it also seems clear that circadian clocks satisfy the conditions of Drestke's criteria for representationhood: they reliably respond to the period and phase of the day-night cycle, and it's presumably in virtue of that responsiveness that circadian clocks have been naturally selected to guide behaviors that depend for their success on being coordinated with the day-night cycle. And again, Dretske (1988, p. 89) himself seems to agree that circadian clocks qualify as representations in his sense.

It's worth noting that in a few animals, circadian clocks do *not* participate in functioning homomorphisms. For example, the Somalian cavefish has evolved for millions of years in the absence of sunlight, and possesses a vestigial circadian 'clock' with a period of 47 h (Cavallari et al. 2011). This mechanism plays in important role in mediating internal metabolic processes, but it is not causally responsive to an external system that it guides behavior with respect to. Thus it doesn't count as a representation in either Gallistel's or Dretske's sense. This example underscores that a 'receptor' condition is crucial for the cogency of a structural notion of representation, just as a 'homomorphism' condition is an essential component of a receptor notion of representation. Again, it seems that whatever satisfies the conditions of one notion satisfies the conditions of the other. There's *simply no distinction* between receptors and structural representations.

One might object at this point that although there's no *categorical* distinction between receptors and structural representations, there's nevertheless a *graded* distinction. Just as there are genuinely bald people, even though the category of baldness is fuzzy at the boundaries, there are genuine receptors, even though that category is fuzzy; receptors are just, as it were, *degenerate* functioning homomorphisms. This is true, but irrelevant. While we might distinguish between very simple and more complex functioning homomorphisms, the question at issue is whether this marks a distinction between representations and non-representations. And I see no theoretical or intuitive reason to think that it does. Ramsey (2007) writes that he's "willing to be fairly unrestrictive about what qualifies as a map, model or simulation, as long as there is a clear explanatory benefit in claiming the system *uses* the structure in question as such" (p. 82). I quite agree; I just think that the notion of a functioning homomorphism captures a clear sense in which a structure is used as a model *even when the homomorphism is very simple*. To reiterate an earlier example, a car's oil light is no less a measuring instrument than a fuel gauge simply because it makes fewer distinctions;

[27] The word I replaced here is 'isomorphism'. In earlier work, Gallistel tended to express his view in terms of isomorphisms, but in more recent work he expresses it in terms of homomorphisms, presumably due to a recognition of the problems with an isomorphism-based view of representation of the kind we discussed earlier.

the richness of the homomorphism it participates in pertains to its representational *expressiveness*, not to its representational *status*.

Ramsey thinks that analogies like these pump our intuitions in misleading directions. He thinks that receptor-like artifacts such as oil lights or pregnancy tests can only function *as* representations when they're interpreted by intelligent agents, and that when we find it intuitive that sub-personal receptors might function as representations in a similar way, "we overlook the fact that when the cognitive agent is removed from the picture. . .the process becomes just like other causal processes that we intuitively judged to be *non*-representational in nature" (p. 218). On the other hand, he thinks that structural representations *can* function as representations without being interpreted by intelligent agents: "[a] mindless system can. . .take advantage of the structural isomorphism between internal structures and the world, and in so doing, employ elements of those internal structures as representations-qua-stand-ins" (p. 200). Note that this objection *presupposes* the distinction I'm objecting to; it's not clear why the quote just mentioned wouldn't hold true of very *simple* 'structural isomorphisms' such as those mediated by binary receptors, as well as the more complex ones that Ramsey seems to have in mind. But merely replying that there is no distinction between intuitively receptor-like and intuitively map-like systems with respect to Ramsey's concerns about expunging the homunculus might be a pyrrhic victory for someone who wants to defend representationalism from Ramsey's arguments, for one might worry that Ramsey's concerns about the automatization of receptor-like surrogates applies to *all* surrogates; as Garzón and Rodriguez (2009) point out, one might worry that the homomorphisms that purportedly underlie a surrogative function can only be fixed by the interpretative faculties of an intelligent agent.

I think this concern is mistargeted; although both Ramsey (2007) and Garzón and Rodriguez (2009) are right to raise concerns about how surrogates could be automated, they fail to identify *when and why* those concerns arise. As I mentioned earlier, when we focus only on structural resemblances that are grounded in the static, 'pictorial' properties of a system, it's difficult to see how those resemblances could be exploited without someone to perceive or interpret them. However, as Craik (1943) insightfully pointed out, if we shift our attention to resemblances that are grounded in the dynamic, functional properties of a system, it becomes clearer how resemblances could be exploited mechanistically; mechanisms can resemble one another not just in how they *look*, but in how they *work*.[28] But that's true regardless of whether the operation of a mechanism sustains simple resemblances or more complex ones. I take it that there's no deep mystery about how, for example, the 'resemblance' between the activity in a fly's thermosensors and the ambient temperature can be exploited to enable the fly to avoid aversive temperatures.

To repeat a familiar refrain, the key question about whether a system counts as a representation concerns not how complex it is, but whether it is *used* as a surrogate

[28] To relate the point here to my earlier characterization of functioning homomorphisms in terms of measurement theory, measurement needn't involve the assignment of *numerals* to a given system, which must be interpreted by an intelligent agent; measurement procedures can be *automated* within a system—think, for example, of how the measurement of temperature is automated within a thermostat. An automated measurement procedure *just is* a procedure that mediates a functioning homomorphism in Gallistel's sense

for another system. This is the central idea underlying Gallistel's 'structural' view of representation, but it's also the central idea underlying Dretske's 'receptor' view. Following *Frank* Ramsey, Dretske frequently employs a metaphor that *William* Ramsey would surely find congenial: representations are *maps by means of which we steer* (*ibid*., p. 79). So Dretske doesn't differ from the latter Ramsey in his conception of what representations essentially *do*, he just has a broader conception of the kinds of mechanisms that can function as internal, behavior-guiding 'maps'. Of course, nobody thinks that maps in the relevant sense need be very much like *cartographic* maps, with contour lines and a compass rose. The arguments in the present section suggest that to exclude receptors from the ambit of a structural conception of representation is to take the metaphor of a cartographic map too literally. Receptors can participate in homomorphisms, and can function as behavior-guiding 'maps' by doing so. Receptors *just are* structural representations.

3.2 Mental representations

At this point, a proponent of structural representations might complain that I've been less than charitable. Surely, the complaint might go, there's *some* distinction to be made between mechanisms like feature detectors and the kinds of mechanisms that psychologists have had in mind when they've talked of 'mental models', which goes beyond a mere difference in the relative richness of the homomorphisms that those mechanisms participate in. After all, mental models are supposed to be *mental*, whereas feature detectors are arguably not that different from simple mechanical devices that might be found in all sorts of non-intentional systems such as thermostats. The picture we get of mental models from pioneering discussions by Craik (1943), and later elaborations by classical cognitive scientists like Johnson-Laird (1983), is that they're introspectively accessible states that agents can manipulate *in thought*, independently of the flux of real-time perception, which enable agents to achieve relatively sophisticated cognitive feats, such as drawing lessons from past experiences, planning for the future, reasoning about counterfactual possibilities, and so forth. This certainly seems to be what Craik (1943) had in mind when he characterized mental models as states that enable an agent to try out "possible actions within its head" (p. 51), and what Johnson-Laird (1983) had in mind when he invoked mental models to explain capacities like counterfactual reasoning. It also seems to be what Cummins (1994) has in mind when he writes that "sophisticated cognition" is enabled by the mind's capacity to "operate on something that has the same structure as the domain it is said to cognize" (pp. 297–298).

The 'sophisticated', stimulus-independent cognitive capacities associated with the picture of mental models I've just sketched have historically engendered the most venerable and compelling arguments for thinking that there are such things as mental representations in the first place. For example, the best argument that Locke provides for believing in *ideas* appeals to our capacity to think about things in memory and imagination. Since those things aren't literally in our minds, Locke claims, "it is necessary that something else, *as a sign or representation* of the thing it considers, should be present to [the mind]; and these are ideas" (Locke 1689 [1975], IV.xxi.4, my emphasis). Clark and Toribio (1994) update this argument for the information-processing age by pointing out that certain cognitive capacities are 'representation-hungry'. As Clark (2001) later puts the point, "Internal representations look prima facie essential for such. . .activities as dreaming of Paris, mulling over U.S. gun control policy, planning next year's vacation, counting the windows of your New York apartment while on holiday in Rome, and so on. . .. All these cases, on the face of it, require the brain to use internal stand-ins for potentially absent, abstract, or non-existent states of affairs" (p. 129).

It seems to me that considerations like these provide a quick argument against Ramsey's anti-representationalist and pro-eliminativist conclusions. It's surely uncontroversial that we *have* stimulus-independent cognitive capacities, such as the capacity to recall episodic memories about specific events in our past, prospectively 'project' ourselves into future episodes, entertain pseudo-perceptual mental imagery, and so forth, and that such capacities stand in need of explanation. Moreover, it seems reasonable to call the mechanisms that explain those capacities, whatever they turn out to be, 'representations', simply by virtue of the capacity they explain—in the same way that it would have been reasonable to call the mechanisms of photosynthesis, whatever they had turned out to be, 'sunlight-to-food converters'.

However, I will not pursue this line further since my primary goal is not to undermine Ramsey's conclusions, but rather to scrutinize the premises that purportedly lead to those conclusions, so as to better understand the explanatory ambit of the widely invoked notion of structural representation. The preceding discussion has suggested that 'sophisticated', stimulus-independent cognitive capacities are mediated by mental representations, and that there's an intuitive distinction between such representations and simple mechanisms like feature detectors, *whether or not* there might be some attenuated sense in which the latter count as representations. But the question at issue is whether that intuitive distinction maps on to the distinction between structural representations and receptors. We just saw that when Clark discusses stimulus-independent cognitive capacities, he emphasizes the importance of appealing to internal *stand-ins*. This seems to accord with Ramsey's emphasis on the surrogative, standing-in role that in his view is constitutive of representationhood, however, the two authors mean quite different things by 'standing-in'. For Clark, a stand-in is a system that can be manipulated *offline*, in the absence of a direct causal connection with the system it stands in for, whereas for Ramsey, a stand-in is a system that enables successful interactions with the system it stands in for, by virtue of the fact that both systems embody the same abstract structure. While the former author doesn't emphasize the structural resemblance aspect, and the latter doesn't emphasize the offline manipulation aspect, nevertheless both aspects seem complementary; a natural thought is that it's precisely because system *A* embodies the same abstract structure as system *B* that *A* can be used as an effective offline surrogate for *B*.[29] So perhaps we might explore the idea

[29] This line of thought *seems* to underlie many discussions of structural representation, which employ terms like 'standing-in' and 'surrogative reasoning'. However, it's often unclear whether these terms are being used in Clark's sense or Ramsey's. My impression is that there's a tendency in the literature to assume that for a representation to be used as a surrogate *just is* for it to be manipulated offline, but as our discussion of Ramsean surrogates has shown, many surrogates are causally coupled to the systems they're surrogates for. Indeed, this is arguably the normal function of surrogates—think of a map being used in conjunction with landmark recognition to navigate through an environment.

that structural representations and receptors are not to be distinguished on the basis of representational *status*, but rather on the basis of the basis of representational *genus*; the idea would be that what demarcates structural representations from receptors is that only the former can be *decoupled* from the systems they're functionally homomorphic with, and moreover that such decoupleable functioning homomorphs are mental models, the vehicles of stimulus-independent cognition. Receptors, *qua* functioning homomorphs, might be representations, but they're not, on this view, *mental* representations.

Now, Ramsey, Gallistel, and many other proponents of structural representations do not *in fact* hold that decoupleability is a necessary condition for something to count as a structural representation, or that structural representations are distinctively mental. Recall one of Gallistel's examples of a structural representation, a circadian clock. Circadian clocks are found not just in animals, but in almost all organisms on Earth, including plants. And like the circadian clocks in animals, those in plants exist for a reason, namely to enable the organism to engage in behaviors (or 'behaviors', if you prefer) that depend for their success on being coordinated with the day-night cycle. So circadian clocks in plants seem to satisfy the conditions for being structural representations; indeed, they function in essentially the same way as those in animals, and involve many of the same evolutionarily-conserved biochemical mechanisms (Kay 1997).[30] Yet arguably they can't generally be manipulated offline, and they're certainly not distinctively mental. In personal communication, Ramsey has agreed that circadian clocks in plants might count as structural representations; he emphasizes that what's essential is whether a system is used as an internal surrogate, not whether it's distinctively mental. But in any case, to see the point at issue, we needn't look to such exotica as circadian clocks in plants. Recall that one of Ramsey's central examples of a structural representation is the S-shaped groove in the toy car depicted in Fig. 1. This cannot in any sense be manipulated 'offline'; it works by being directly coupled to the car's motion. Moreover Ramsey expressly discusses this example to make the point that structural representations aren't distinctively mental; he argues that structural representations can perform their surrogative function in mindless systems like toy cars, and hence can function *as* representations without being interpreted by an intelligent homunculus (see pp. 193–203).

Nevertheless, we might take the proposal that structural representations are necessarily decoupleable as a friendly amendment to Ramsey's view, which strengthens his anti-representationalist and pro-eliminativist argument. For it seems to provide a way for Ramsey to distinguish structural representations from receptors, and hence to avoid the skeptical argument I developed in the Sect. 3.1. Moreover it does so in a way that is congenial to the eliminativist strand of his argument. Ramsey's eliminativist argument, as it stands, seems strangely at odds with his view of representation. Recall the contours of the argument: while classicism once attempted to explain cognition by invoking genuinely representational 'structural' states, connectionism is eclipsing classicism and promises to explain cognition in terms of non-representational 'receptors'. But what

[30] To underscore the point made in the previous section, note also that circadian clocks in plants satisfy the conditions for being Dretskean receptors, for essentially the same reasons.

if connectionist explanations were to prove inadequate, and classicism experienced a resurgence? If, as Ramsey seems to think, there's nothing distinctively *mental* about the structural representations posited by classical explanations, how would the explanatory success of classicism vindicate intentional psychology and dispel the bogey of eliminativism? Whether or not a cognitive theory threatens the existence of mental states seems to turn on whether it posits mental representations, not representations *simpliciter*.[31]

But while it might be charitable to stipulate that decoupleability is a necessary condition for something to be a structural representation, I see no corresponding reason to hold that the *lack* of decoubleability is a necessary condition for something to be a receptor. Many of the most representative models of the neo-connectionist literature of the '80s were simulations of stimulus-independent cognitive capacities, and didn't assume that the representations they modeled could be elicited only by immediate perceptual stimulation. Indeed, in the 'bible' of neo-connectionism, Rumelhart et al. (1986) explicitly address the question of how activity in neural networks might be manipulated offline in the course of cognitive processes like planning or counterfactual reasoning, and suggest that this is achieved by a "mental model" instantiated by a "relaxation network which takes as input some specification of the actions we intend to carry out and produces an interpretation of 'what would happen if we did that'" (p. 41), thereby generating predictions about events in the world. In accordance with this general picture, recent evidence suggests that even feature detectors can be elicited offline; for example, neurons in early visual cortex seem to be activated during visual mental imagery (Kosslyn and Thompson 2003). Of course, one might *stipulate* that receptors cannot be manipulated offline so as to preserve the distinction between structural representations and receptors, but the resulting notion of a receptor wouldn't be the one that could feature in Ramsey's anti-representationalist argument, for there's no reason to think that connectionists employ such an etiolated notion.

Putting Ramsey's argument aside, haven't we at least found that by cashing the notion of structural representation in terms of decoupleable functioning homomorphs, we've helped to elucidate the nature of 'mental models', the vehicles of stimulus-independent cognition? Not at all. For despite what I suggested earlier, circadian clocks in plants *can* be manipulated offline in the service of anticipatory behavior. In a fascinating discussion of adaptive behavior in plants, Garzón and Keijzer (2011) point out that some plants, notably *Lavatera cretica*, are capable of reorienting their leaves overnight so as to face the 'anticipated' location of the sun in the morning, thereby maximizing daily light intake. The circadian clocks in *Lavatera* can sustain this behavior over several days in the absence of sunlight (Schwartz and Koller 1986). Similarly, the circadian clocks in plants of the genus *Arabidopsis* help to muster

[31] Of course, if one could show that a cognitive theory doesn't posit representations in *any* sense, one would have shown that it doesn't posit *mental* representations. But as I pointed out at the beginning of Sect. 3, Ramsey's argument doesn't show that; even granting all the premises, the most the argument could show is that connectionist mechanisms aren't representations *by virtue of functioning as receptors*, not that such mechanisms fail to be representations *simpliciter*. Indeed, the naturalistic methodology that Ramsey adopts precludes him from developing the kind of argument at issue; recall that Ramsey rightly holds that it is forlorn to seek a general analysis of representation that will encompass all and only representations.

an anticipatory chemical defense against herbivorous insects overnight, even in the absence of sunlight (Goodspeed et al. 2012). The circadian clocks in these plants count as functioning homomorphs, as I argued earlier, and there seems to be a robust and interesting sense in which they can be used offline, so they seem to count as structural representations in the revised sense under consideration; nevertheless, they surely don't count as *mental* representations, i.e. the vehicles of cognitive processes like episodic memory, planning, and mental imagery. Although Garzón and Keijzer (2011) take the evidence of offline anticipation in plants to show that plants are "cognitive in a minimal, embodied sense" (p. 166), they don't hold that plants have minds in the full-blooded sense that's presently at issue. They rightly point out that for the purpose of understanding the continuity between life and mind, it is important to explore the oft-neglected commonalities between plant and animal behavior, but that "it goes without saying that for other purposes the emphasis may be rightly placed upon the [differences]" (p. 156). Just so; for the purpose of elucidating the nature of mental representations, the notion of structural representation—even if supplemented with a decoupleability condition—simply won't do.[32] The claim that structural representations "make sophisticated cognition possible" might be true, but only in the sense in which *atoms* make sophisticated cognition possible; it can't be the manipulation of structural representations *as such* that makes sophisticated cognition possible, since structural representations are manipulated in all sorts of mindless systems, such as plants.

4 Toward a new job description

In attempting to articulate a notion of representation that will play a role in cognitive explanations, philosophers and psychologists have frequently appealed to notions like isomorphism and structural resemblance. Proponents of these 'structural' notions of representation often develop their views against the backdrop of the picture of mental models that I sketched at the beginning of the previous section: states that can be manipulated in thought, during cognitive processes like episodic memory, counterfactual reasoning, and mental imagery. However, when these theorists attempt to articulate what exactly structural representations *are*, the background picture tends to fall away, unnoticed. We see this clearly in Ramsey's (2007) discussion of structural representations. Although Ramsey often characterizes structural representations as 'mental models' of the kind discussed by Johnson-Laird (1983), we've seen that the substance of his view is best captured by Gallistel's (1998) notion of a *functioning homomorphism*, and that there's nothing distinctively *mental* about representations so understood—even if we add the further condition that functioning homomorphs must be *decoupleable* from the systems that they're homomorphic with. Functioning

[32] I should note that I don't advocate an anthropocentric view of the domain of organisms that possess mental representations. I think that many animals, and probably even insects, have mental representations. For example, Clayton et al. (2001) have provided compelling evidence that scrub jays have episodic memory. But the ingenuity of this experiment, and the controversy surrounding its interpretation, highlights the fact that the capacities that mental representations mediate are highly non-trivial to demonstrate, and are substantially different from the 'cognitive' capacities of plants.

homomorphs, whether decoupleable or not, can be found in all sorts of non-mental systems, such as plants. Indeed, Ramsey explicitly endorses the claim that structural representations can be found in non-mental systems, as a premise in his argument that such representations needn't be interpreted by a homunculus to function *as* representations.[33]

The widespread tendency to conflate structural representations with mental models I think helps to explain some of the intuitive appeal of Ramsey's argument against the representational *bona fides* of connectionist mechanisms. There seems to be a clear difference between the mental models mediating our thoughts about the past and future, and, say, single neurons that respond to oriented lines. So if one identified this difference with the purported distinction between structural representations and receptors, one might find Ramsey's argument compelling. But this would be a threefold mistake. First, it would rest on an etiolated caricature of the explanatory repertoire of connectionists; mainstream connectionists appeal to a range of complex, hierarchically structured systems to explain 'sophisticated' cognition. Second, it would presuppose a distinction between structural representations and receptors that simply has no theoretical content. And third, it would assume, falsely, that structural representations are distinctively mental.

In closing, I'd like to look at this third assumption more closely, since I think it can help diagnose where Ramsey's reasoning went awry. Although Ramsey officially holds that a representation is any internal mechanism that functions as a behavior-guiding map, his thinking sometimes seems to be guided by a tacit conception of representation as something distinctively mentalistic. When he evaluates the representational credentials of structural and receptor representations, he seems to vacillate between these two conceptions. He seems to evaluate receptor representations with respect to a mentalistic conception of representation, and rightly concludes that they're not representations in this sense, yet he seems to evaluate structural representations with respect to a non-mentalistic conception of representation, and rightly concludes that they *are* representations in this sense. But because of the tacit equivocation, he fails to notice that structural representations and receptors are the same in both respects: both are representations in that they function as behavior-guiding maps, but neither plays a distinctively mentalistic role in doing so.

This tacit equivocation sometimes becomes explicit. Recall that Ramsey sometimes characterizes the role of an internal, behavior-guiding map in terms of what Swoyer (1991) calls 'surrogative reasoning'. Consonant with a strict interpretation of 'surrogative reasoning', as something that only *agents* are capable of, Ramsey holds that the central difference between receptors and 'real' representations is that the latter, but not the former, participate in "a process that is properly or naturally viewed as something

[33] Incidentally, this strategy for showing that structural representations can function as representations in the absence of a homunculus—what Ramsey calls the 'mindless strategy' of showing that structural representations can play a role within non-mentalistic systems, and hence aren't distinctively mental—seems to conflate two distinct questions. It's one thing to ask whether a type of representation is distinctively mental, and it's quite another to ask whether a type of representation can function as a representation within a purely mechanistic system. While the mindless strategy might be *sufficient* for expunging the homunculus, it's surely not *necessary*; to suppose otherwise would seem to assume a kind of dualism according to which minds cannot be explained mechanistically.

like *learning about* or *making inferences about* some state of affairs with which the representation stands in some nomic relation" (2007, p. 141). However, we've also seen that Ramsey often thinks of surrogative reasoning far more liberally, as a process that, for instance, might be carried out by the mechanically-driven rudder in the toy car depicted in Fig. 1. It's this liberal conception that underwrites Ramsey's claim that the "surrogative, 'standing-in-for' property is *not* dependent upon any inference or learning process" (*ibid*., p. 201, my emphasis). So Ramsey seems to think of surrogative reasoning as a distinctively *mentalistic* process when evaluating the representational status of receptors, and denies that receptors are representations on that basis, yet he thinks of it far more liberally when evaluating structural representations.

I think this reveals two important lessons for discussions about the role of representation in cognitive explanations. The first is that it's crucial to be clear about whether or not 'representation' is being used as shorthand for 'mental representation'. Just as there's little reason to think that all the things we call representations, ranging from traffic lights to thermoreceptors, fall under the same explanatory generalizations, we shouldn't assume that sub-personal representations and mental representations have much in common beside the label 'representation'. The second is that, contrary to widespread assumptions, the notion of a structural representation, *qua* an offline, behavior-guiding map, doesn't help to explicate the distinctive nature of mental representation. One might complain that in describing my purported counterexample of a circadian clock that is used offline to control anticipatory leaf reorientation behavior in plants, I'm using terms like 'offline control', 'anticipation' and 'behavior' in ways that are far more liberal than proponents of a mentalistic construal of structural representations would like. But that's precisely my point; simply invoking those terms isn't enough to capture a notion of distinctively mental representation. One must articulate a more robust interpretation of those terms.

How then do we capture a notion of distinctively mental representation? That's a story for another time. However, I'd like to close with a couple of suggestive observations. Contrary to some caricatures of connectionism as being an heir to behaviorism, research into the mechanisms of 'representation hungry', stimulus-independent cognitive capacities is one of the most active and productive areas of contemporary neuroscience. The emerging picture, supported by a wealth of neuroimaging, neuropsychological, and neurophysiological evidence, is that capacities like episodic memory, prospective 'mental time travel', and mental imagery involve overlapping functional networks with a common core, which subserves a simulative function much like that envisioned by Rumelhart et al. (1986): top–down signals elicit a 're-enactment' of activity in cortical areas involved in bottom-up perception.[34] These areas contain feature detectors of the kind that Ramsey disparages as mere 'receptors'. So, far from undermining our self-conception as intentional agents with rich mental lives, contemporary connectionism seems well on its way to revealing the mechanisms of our mentality. And the humble feature detector seems to be playing a crucial role in that endeavor.

[34] See Danker and Anderson (2010), Kent and Lamberts (2008), and Schacter et al. (2008) for reviews.

Acknowledgments Special thanks to Bill Ramsey for writing such a stimulating book and for generous discussion. Thanks also to Paco Calvo, Frances Egan, Bob Matthews, Lisa Miracchi, Gualtiero Piccinini, Ron Planer, and an anonymous referee for helpful comments.

References

Allen, C., & Hauser, M. (1993). Communication and cognition: Is information the connection? *Philosophy of Science*, *2*(8), 81–91.

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

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

Cavallari, N., Frigato, E., Vallone, D., Fröhlich, N., Lopez-Olmeda, J., Foà, A., et al. (2011). A blind circadian clock in cavefish reveals that opsins mediate peripheral clock photoreception. *PLoS Biology*, *9*(9), e1001142.

Churchland, P., & Churchland, P. (2002). Neural worlds and real worlds. *Nature Reviews Neuroscience*, *3*, 903–907.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language*, *16*(2), 121–145.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, *101*(3), 401–431.

Clayton, N., Yu, K., & Dickinson, A. (2001). Scrub jays (*Aphelocoma coerulescens*) form integrated memories of the multiple features of caching episodes. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(1), 17–29.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.

Cummins, R. (1994). Interpretational semantics. In S. Stich & T. Warfield (Eds.), *Mental representation: A reader* (pp. 297–298). Cambridge, MA: Blackwell.

Danker, J., & Anderson, J. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87.

Dennett, D. (1981). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, *3*(1), 1–8.

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

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, *17*(6), 1276–1314.

Fodor, J. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind*, *94*(373), 76–100.

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. (1998). Symbolic processes in the brain: The case of insect navigation. In D. Osherson, D. Scarborough, L. Gleitman, & D. Sternberg (Eds.), *An invitation to cognitive science: Methods, models, and conceptual issues* (2nd edn., Vol. 4, pp. 1–52). Cambridge, MA: The MIT Press.

Gallistel, C., & King, A. (2010). *Memory and the computational brain*. Oxford: Wiley-Blackwell.

Garzón, F., & Keijzer, F. (2011). Plants: Adaptive behavior, root-brains, and minimal cognition. *Adaptive Behavior*, *19*(3), 155–171.

Garzón, F., & Rodriguez, A. (2009). Where is cognitive science heading? *Minds and Machines*, *19*(3), 301–318.

Godfrey-Smith, P. (2006). Mental representation, naturalism, and teleosemantics. In D. Papineau (Ed.), *Teleosemantics: New philosophical essays* (pp. 42–68). Oxford: Oxford University Press.

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. Indianapolis, IN: Bobbs-Merrill.

Goodspeed, D., Chehab, E., Min-Venditti, A., Braam, J., & Covington, M. (2012). Arabidopsis synchronizes jasmonate-mediated defense with insect circadian behavior. *Proceedings of the National Academy of Sciences o the United States of America*, *109*(12), 4674–4677.

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*. http://ndpr.nd.edu/news/23327-representation-reconsidered/.

Hamada, F., Rosenzweig, M., Kang, K., Pulver, S., Ghezzi, A., Jegla, T., et al. (2008). An internal thermal sensor controlling temperature preference in drosophila. *Nature*, *454*(7201), 217–220.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Isaac, A. (2012). Objective similarity and mental representation. *Australasian Journal of Philosophy*, 1–22.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kay, S. (1997). PAS, present and future: Clues to the origins of circadian clocks. *Science*, *276*(5313), 753–754.

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: Retrieval as mental simulation. *Trends in Cognitive Sciences*, *12*(3), 92–98.

Kirsh, D. (1990). When is information explicitly represented. In P. Hanson (Ed.), *Information, language and cognition* (pp. 340–365). Vancouver: University of British Columbia Press.

Kosslyn, S., & Thompson, W. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin*, *129*(5), 723–746.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (Eds.). (1971). *The foundations of measurement*. New York: Academic Press.

Lehky, S., & Sejnowski, T. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature*, *333*(6172), 452–454.

Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, *47*(11), 1940–1951.

Lewis, D. (1971). Analog and digital. *Nous*, *5*(3), 321–327.

Locke, J. (1689 [1975]). *An essay concerning human understanding*. Oxford: Oxford University Press.

Miall, C., & Wolpert, D. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279.

Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

O'Brien, G., & Opie, J. (2001). Connectionist vehicles, structural resemblance, and the phenomenal mind. *Communication and Cognition*, *34*, 1–2.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. Bloom-Lloyd (Eds.), *Cognition and categorization* (pp. 259–303). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rumelhart, D., Smolensky, P., Mcclelland, J., & Hinton, G. (1986). Schemata and sequential thought processes in pdp models. In J. McClelland, D. Rumelhart, & the PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models, Chap. 14 (pp. 7–57). Cambridge, MA: MIT Press.

Schacter, D., Addis, D., & Buckner, R. (2008). Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences*, *1124*, 39–60.

Schwartz, A., & Koller, D. (1986). Diurnal phototropism in solar tracking leaves of *Lavatera cretica*. *Plant Physiology*, *80*(3), 778–781.

Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science*, *63*(3), 519–545.

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 William Ramsey, 'representation reconsidered'. *The British Journal for the Philosophy of Science*, *62*(3), 669–675.

Sterelny, K. (1995). Basic minds. *Philosophical Perspectives*, *9*, 251–270.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449.

van Fraassen, B. (2008). *Scientific Representation*. Oxford: Oxford University Press.