Routledge
Taylor & Francis Group

Philosophical Explorations

An International Journal for the Philosophy of Mind and Action

ISSN: 1386-9795 (Print) 1741-5918 (Online) Journal homepage: http://www.tandfonline.com/loi/rpex20

Representation and mental representation

Robert D. Rupert

To cite this article: Robert D. Rupert (2018) Representation and mental representation, Philosophical Explorations, 21:2, 204-225, DOI: 10.1080/13869795.2018.1477979

To link to this article: https://doi.org/10.1080/13869795.2018.1477979

Published online: 02 Jul 2018.

Philosophical Explorations, 2018
Vol. 21, No. 2, 204–225, https://doi.org/10.1080/13869795.2018.1477979
This paper is part of the special issue "Enactivism, Representationalism, and Predictive Processing" guest edited by Krzysztof Dołęga, Luke Roelofs, and Tobias Schlicht.

Representation and mental representation

Robert D. Rupert*

Department of Philosophy and Institute of Cognitive Science, University of Colorado at Boulder, Boulder, CO, USA

(*Received 4 May 2018; final version received 4 May 2018*)

This paper engages critically with anti-representationalist arguments pressed by prominent enactivists and their allies. The arguments in question are meant to show that the "as-such" and "job-description" problems constitute insurmountable challenges to causal-informational theories of mental content. In response to these challenges, a positive account of what makes a physical or computational structure a mental representation is proposed; the positive account is inspired partly by Dretske's views about content and partly by the role of mental representations in contemporary cognitive scientific modeling.

Keywords: representation; mental representation; enactivism; cognitive science

I. Introduction

Use of "representation" pervades the literature in cognitive science. But, do representations actually play a role in cognitive-scientific explanation, or is such talk merely colorful commentary, something to spice up the introduction and discussion sections of scientific papers? Are, for instance, patterns of cortical activity in middle temporal visual area or strings of symbols in a language-processing parser genuine representations (of motion and of syntactic structure, respectively)? Do they have content? And if they do, can a naturalist assign such contents in a well-motivated and satisfying way?

Enactivists have generally been hostile to, or at least uninterested in developing, representational accounts of cognition (Chemero 2009; Stewart, Gapenne, and Di Paolo 2010), and they tend to answer the preceding questions accordingly. Dan Hutto and Erik Myin (2012), for example, have forcefully criticized naturalistic theories of mental representation of the sort that seem to find a comfortable home in the philosophical foundations of cognitive science, arguing that such naturalistic theories founder on what they call the "hard problem of content" (Hutto and Myin 2012, chapter 4). Although the field teems with naturalistic proposals, in no case, according to Hutto and Myin, does a naturalistic theory assign contents that play the appropriate role. The structures in question do not contribute as such, that is, *as representations*; they do not contribute in virtue of their fulfilling the job description of representations (Hutto and Myin 2012, 61–62). William Ramsey

*Email: robert.rupert@colorado.edu

expresses similar criticisms of at least some naturalistic approaches to content, particularly with regard to the as-such and job-description problems (2007).

The present essay focuses on causal, informational, nomic, and covariation-based theories of content (henceforth grouped together as "causal-informational" theories): theories that take the fundamental component of the content-determining relation to be a distinctive kind of causal pattern or law-based regularity that connects a mental representation to what it represents. According to Ramsey and Hutto and Myin, causal-informational theories fail by the naturalist's own lights; they do not deliver content-bearing units that play an appropriate explanatory role, either because the units in question do not play a genuine representational role at all (possibly by failing to have genuine content) or because, even if they do seem to play a representational role, they do not participate in cognitive processing *qua representations*, or *as such*, that is, in virtue of the features that qualify them as representations.

In this paper, I defend causal-informational theories against such attacks, arguing that when these theories are set in the proper context – that of the cognitive-scientific modeling of intelligent behavior – they can deliver *mental* representations that play an explanatory role as such, regardless of whether they play this role *qua* representations *simpliciter*. In fact, I shall have little to say here about representation *simpliciter*. The proposal made herein, concerning mental representation, is consistent with both an eliminativism about the genus *representation* and with a pan-representationalism. I take no stand on this matter. Rather I characterize only a conception of the kind *mental representation*, in a way that sits well with causal-informational theories of content.[1] Regarding pan-representationalism, causal-informational theories tend to ground representation in such ubiquitous relations as indication and detection – the obtaining of one state in the world's indicating, perhaps by high statistical correlation, that some other state obtains – or law-like dependencies between the occurrence of different states of the world. Such relations permeate nature, which might give some readers pause. Do causal-informational views of mental content entail that representations permeate nature? No, they don't, but arguing that point is not my primary purpose. The matter at hand is the nature of mental representation.

Irrespective of whether detectors, indicators, and various nomologically connected causal *relata* count as representations outside of the mental or cognitive context, I propose that they become *mental representations*, functioning *as such*, when the three following conditions are met: (a) they appear in properly mental systems, that is, when they play an explanatory role in an architecture the interaction of various components of which accounts for the *explananda* of cognitive science, namely intelligent behavior; (b) their contribution to the guidance of such behavior rests partly on their representation-like capacities – e.g. that they differentially contribute to the production of behavior concerning the very things that our best causal-informational theory of content assigns as their content; and (c) their playing, or coming to have, the explanatory role in question depends on the presence in such architectures of distinctively cognitive forms of processing, such operations as backpropagation of error (Rumelhart, Hinton, and Williams 1986) or the formation of a single (perhaps more flexibly applicable) production rule as the result of past uses of the same linked chains of production rules to solve similar problems (Rosenbloom et al. 1991). A mental representation contributes its basic (perhaps proto-) representational capacity to the modeling of intelligent behavior, by appearing in complex structures the distinctive operations on which harness that basic (or proto-) representational capacity for the production of such behavior. The satisfaction of these conditions distinguishes mental representations from nonmental representations, regardless of whether what I've labeled their "basic (or proto-) representational capacity" is genuinely representational, merely proto-

representational, or is not deeply representational in any way – because, say, the distinction representational–nonrepresentational does not carve nature at its joints.

The positive proposal summarized in (a), (b), and (c) adapts some of the central ideas of Dretske (1981, 1988, 1997) to which, in my view, Hutto and Myin and Ramsey give short shrift. Thus, after preliminaries in Section II, Section III lays out the essentials of Dretske's framework. Section IV articulates my positive proposal, and Section V brings the proposal, and Dretske's work, into closer contact with the critics' concerns. Section VI closes by offering diagnoses of the disagreement.

II. Remarks about naturalistic methodology

The most well-known and influential naturalistic theories of mental content (Dretske 1981, 1988; Fodor 1987, 1990; Millikan 1984; Papineau 1984) are offered largely as attempts to ground folk psychology – our everyday, belief–desire–intention-based explanations of behavior – in the natural order. They took such a target for good reason. Fodor (1987) emphasizes the apparent value of folk-psychological thinking for such purposes as predicting and explaining the behavior of others and effecting social coordination. Given such success, a naturalist will want to know how to fit the posits of folk psychology – psychological states, entities, and laws, as characterized by the folk – into the natural world. By Fodor's lights, such a goal can be met by demonstrating the consilience of folk psychology with both our emerging cognitive science and our general understanding of the physical world.[2]

Other philosophers of Fodor's generation were less impressed by the causal-explanatory and predictive performance of folk psychology and, at least with respect to some folk-psychological constructs, recommended their elimination from our set of metaphysically serious ontological commitments (Churchland 1981). I do not press such a hard line, but I do recommend pursuit of the relevant science, without regard for potentially constraining commitments of folk psychology, to see what it delivers. Of course, our best cognitive science had better, in the end, account for whatever success folk psychology has, but at present we do not yet know either the extent of folk psychology's success or what best accounts for whatever success folk psychology has – whether, for example, cognitive science will vindicate folk psychology's posits (as Fodor predicted it would) or, rather, recommend seeing folk psychology as some kind of fiction, and provide an error theory to explain why at least some of the component claims of folk psychology can be useful even though false.

It is no surprise, then, that a contrasting set of naturalistic theories of content approaches the topic as something closer to a question in the philosophy of science. What might reasonably be called the "second generation" of naturalistic theorists of mental content – Rupert (1998, 1999), Usher (2001), Prinz (2002), Eliasmith (2003), Ryder (2004), Shea (2007, 2014) – have, on the whole, been inclined to pursue a theory of representational content as part of the theoretical foundations of cognitive science, putting folk psychology largely to one side. In this potted history, Rob Cummins (1989, 1996) plays the trailblazer, insisting that we see the representation-relation(s) itself as a theoretical posit, any particular account of which is to be evaluated by the success of the cognitive-scientific work it underpins.

This distinction makes an enormous methodological difference. Second-generation naturalists treat content in the way theoretical scientists treat any property, kind, or relation. They are free to *stipulate* how and under what conditions it appears, and then run the empirical-scientific gauntlet.[3] The proof is in the pudding, in the contribution the proposed content-fixing relation makes to the causal-explanatory enterprise in question. Thus, it is no

objection to a second-generation causal-informational theory of representational content that the theory fails to retain some particular aspect of the folk conception or that one cannot see how *that* (the relation stipulated by the naturalistic theory of content in question) determines content[4] – unless, of course, the omission at issue seems to correlate with an empirical failure of the theory of content in question (or more likely of the broader enterprise it is a part of), that is, a failure to account for observable, measurable data in the domain in question.[5]

Hutto and Myin, for example, baldly assert, "Bits of the world do not indicate other bits of the world" (2012, 183), as an objection to Dretske's treatment of indication as the material from which representation is wrought. But, as an internal criticism of a naturalistic theory, Hutto and Myin's pronouncement falls flat. When a naturalist makes a model of the content-fixing process, whether it is to ground some folk-psychological truth or to account for the measurable, replicable data, the naturalist gets to stipulate what counts as indication (which Dretske does mathematically, more on which in Section III). If a model that makes use of the stipulated relation satisfies whatever cross-level *desiderata* the first-generation theorist has in mind or, in the case of the second-generation theorist, does the right sort of work in modeling measurable aspects of intelligent behavior, it makes no difference that the terms used to describe the model do not refer to states or processes with all of the properties we are inclined, from the armchair, to associate with those terms.[6]

Stipulation obviously has its limits, though. We should wonder how much the characterization of a property, kind, or relation, posited as part of the cognitive-scientific enterprise, can differ from the pretheoretical conception and remain the thing so-called (Stich 1996, chapter 1). This puzzle has no easy solution. A naturalist will look at the actual scientific history and see messy terrain; in our best science, the continued use of a term across changes in some of the properties associated with it tends to be negotiated on the ground, rather than by adverting to *a priori* theoretical principles governing the stability of reference (Wilson 2008). Much of what follows should be read as a contribution to that process of negotiation, one that favors the retention of "mental representation" for detector-style representations, that is, cognitively unstructured units subject to a causal-informational semantics.

III. The Dretskean framework

In *Explaining Behavior: Reasons in a World of Causes* (Dretske 1988), Dretske develops a theory of representational content that builds, at least indirectly, on his earlier book *Knowledge and the Flow of Information* (1981). The title of the former work makes Dretske's overarching goal clear enough: to explain, within a naturalistic framework, how reasons can function as causes, thus explaining the behavior they rationalize. His theory of content does much of the work in this regard, tying the conditions in which representational content is fixed to the conditions in which content, so fixed, takes on causal relevance (Dretske 1988, 79–80).

With regard to content itself, the apparent problem, as Dretske see it, is this: Representations, as token physical states (syntactially individuated vehicles, neural structures, symbols in a language of thought – whatever), clearly can act as "triggering causes" (as he calls them [1988, 42], much like what Ramsey refers to as mere causal mediators [2007, 135; see also 2016, 8; and Hutto and Myin 2012, 62, 81]). But, that alone does not show that they have their effects in virtue of their semantic properties, as opposed to, say, their physical properties. An opera singer might produce sounds that have semantic content and that also break glass; but the sound waves' breaking of the glass may have

nothing whatever to do with the semantic content of the words sung. Making matters worse, in many cases, content appears to be relational, the representational state being about some other part of the world, which could be well removed from the representational vehicle in question. In contrast, triggering (or mediating) causes are normally thought to do their work locally.

How *can* content make any difference at all, then? Dretske appeals to three central notions: (a) that of information (or indication – these are equivalent; Dretske 1988, 58–59), (b) that of reinforcement learning (Dretske 1988, chapter 5), and (c) that of a structuring cause. Let us begin with Dretske's notion of information. The activation of cognitive unit *v* carries the information that *i*, where *i* is whatever collection of possible states of the environment (transmitter states, as Dretske sometimes calls them) is such that $P(i|v) = 1$. In other words, *v* carries the information that some state in the (possibly gerrymandered) region *i* of the space of relevant events obtains, where *i* is the minimal collection of states of the world such that it is guaranteed that at least one among such states obtains when *v* is activated, given the laws of nature and "channel conditions" (which will be set aside); in other words, by stipulation, *v* carries the information *i* for just whatever *i* is such that the conditional probability of *i* obtaining given the activation of *v* is one. Thus, to use an informal example, if the subject sometimes mistakes horses for cows, then the information carried by the activation of the cognitive unit "cow" should be thought of disjunctively – as "cow being present or horse being present." And, if other things can be mistaken for a cow, then they must be added to the disjunction, that is, to the collection of event types that constitutes *i*. (I describe *i* as a disjunctive list of possible conditions for activation of *v*, but *i* should be thought of more technically, within the framework of probability theory, as an event that corresponds to a region of the sample space that might include more than one outcome, using "outcome" in its technical sense).

Dretske's approach to indication or information does not allow for misindication or misinformation (1988, 56, 65). Thus, on Dretske's view, merely indicating or information-carrying states are not representations; for the ability to misrepresent – to get things wrong – is central to our concept of representation, if anything is (1988, 64, 65).

Enter reinforcement learning. According to Dretske, an information-bearing unit becomes a representation when it gains control (or maintains control that was innately determined in the first instance) of a kind of movement in the organism and such gaining of control results from reinforcement signals, themselves resulting from the behavioral success in a case in which the information-bearing unit caused the movement that led to success:

> It is only when we get to a form of learning whose success depends on the deployment and the use of internal indicators that it becomes plausible to think that the causal processes constitutive of behavior may actually be explained by facts about what these indicators indicate. (ibid. 96)

This process turns mere indication into representation, and in the typical case, the content of the representation amounts to only a proper subportion of what the indicator indicates. Events of reinforcement are token events (or processes). On any particular occasion of the activation of *v*, *v* might or might not contribute causally to the production of a behavioral response. If it does contribute, and the behavior is successful (where success is cashed out by Dretske in terms of the reinforcement of the connection between the indicator and the production of the behavior in question), then it is possible to differentiate between the different component outcomes in *i*. Was it, for example, in virtue of *v*'s copresence, on that occasion, with a cow that the action was successful, or was it in virtue of *v*'s copresence with a horse that the action

in question was successful? If, in the actual learning history of the organism, cows, not horses (or anything else), acted as a reinforcing stimulus (so to speak) for *v*, then, even though *v*, in fact, indicates "either cow or horse or … ," *v* has acquired the function of indicating cows; it has become a representation of cows, even though P(cow present | tokening of *v*) < 1 – even though the activation of *v* is still sensitive to the presence of horses and thus does not actually carry the information that a cow is present (because, by stipulation, indication and information-carrying require a conditional probability of one relative to what's indicated or what the structure in question carries information about).

The process of reinforcement learning allows Dretske to solve his original problem in an indirect way. The content of a mental representation does not function as an internal triggering cause of movement, Dretske admits; the proximate cause of behavior is always a physical state, the physical properties of which screen off the relation to what's represented. Nevertheless, according to Dretske, content has causal-explanatory relevance because it can act as a structuring cause. Actions are processes; they have a causal structure extended in time. An action consists of *this* event following *that* event and leading to this *other* event. Structuring causes are not links in such a causal chain. Rather, they are causal accounts of why one kind of causal chain is in place in the subject as opposed to some other causal chain's being in place, that is, accounts of the appearance of a process with *this* causal contour rather than some *other* causal contour. It is not *v*'s representing cows that causes the subject's movement across the barn to, for instance, retrieve a milking stool, when *v* is active. But, the reinforcement-based process that gave *v* its content ("cow" rather than "cow or horse or … ") explains why *v* is, in this subject, rigged up to the behavior of retrieving a milking stool, rather than being rigged up to the process of fetching a saddle. There being such a process in place – such a structured series of causes – results from the reinforced priority, in the subject's learning history, of one of the elements in *i* over the others. Coming to have content "cow" in our example just is the process of having *v* rigged up to certain forms of behavior; *v*'s sensitivity to cows explains why that rigging came to exist (in token instances in which *v* co-varied with a horse, and caused a milking stool was fetched, the connection between *v* and the fetching of a milking stool was not reinforced). Representational content plays an explanatory role as a structuring cause (it is responsible for the shape of the causal processes that appear in the subject), even though it is not a triggering cause.[7]

IV. How representation-like structures become mental representations

In this section, I build on some of Dretske's central ideas, while also departing from his view in substantial ways.[8] I aim to account for the role of representations, not in folk-psychological explanations of behavior (as the products of belief–desire pairs), but rather in cognitive-scientific explanations of behavior.[9] Moreover, in previous work, I criticize the details of Dretske's content-fixing proposal and attempt to improve on it (Rupert 1999, 2008a; see also Slater 1994). These departures duly noted, Dretske's approach inspires significant aspects of what follows, in particular, my emphasis on the role of the *explanandum* and on the distinction between representations and *mental* representations.[10]

On the proposed view, for a unit to be a *mental* (or cognitive) representation, it is not enough that it play some explanatory role or other in virtue of the information it carries or in virtue of its other covariation-related properties; nor is it enough that the unit has acquired a control function in a system because of its causal-informational properties. Rather, being a mental representation is a matter of being a representation that contributes

its covariation-related (or structural-iconic)[11] properties to the functioning of a system the activities of which account for the distinctive *explananda* of cognitive science by the implementation or running of processes distinctive to cognitive-scientific modeling.

Cognitive science has its proprietary set of data and distinctive models of those data. Like any science, it begins with an educated guess, that a certain range of data will ultimately be modeled using a relatively unified set of tools that contrast with those used to model data in other domains. In cognitive science, the data in question, at least at a first pass, involve language use, theory construction, proof production, the coordination of the individual's behavior with that of others, the products of art and architecture, the alteration of one's behavior in nonrandom correlation with the structure of objects that the subject has interacted with in the past, highly structured behavior such as the playing of games, and so on – all of the behavior we might pretheoretically place under the rubric of intelligent behavior. And, as with any science, as cognitive science has developed, the manner of collecting data has become more focused. Controlled experiments plumb differences in outputs in highly circumscribed contexts, with carefully measured independent and dependent variables.

Of course, as with any science, the guesses of cognitive scientists can turn out to be wrong. Our best models of some of the data might not be of a piece with our best models of other bits of the data; as a result, the set of target data – those thought to be proprietary *explananda* of the science – might be reduced in scope, some of them now excluded from the domain of cognitive-scientific *explananda*. Or, in some cases, the cognitive-scientific enterprise itself might bifurcate, deep differences in kinds of successful models suggesting that the *explananda* reflect two very different kinds of phenomena after all. Alternatively, when a significant portion of the data have been successfully modeled within a relatively unified framework, cognitive scientists might then discover that the same sorts of models account for kinds of behavior that were not initially thought to be among the *explananda* of cognitive science, in which case the scope of cognitive science expands. Having acknowledged the fluid nature of the enterprise, consider three aspects of cognitive-scientific work that have been, and remain, central to it.

First, consider the integral role played by the idea of an architecture or a cognitive system (Rupert 2004, 2008b, 2009, 2010, 2013). The components and structure of the architecture vary, depending on the style of modeling in question (computational, connectionist, subsumption-based, brute neural, dynamicist, and so on). But, the point of present importance is that a detector- or indicator-based representation contributes to the explanation of behavior in the context of the architecture; that is the typical context for the contribution of any element of the model, besides the data themselves.

Second, as noted above, the cognitive-scientific enterprise aims to account for a distinctive range of observable, measurable data: roughly speaking, those to do with intelligent behavior. Combining this point with the point about architectures, I mean to emphasize the role of mental representations as contributors to distinctive styles of modeling (styles not found, for example, in models of the immune system [Ramsey 2007, 125, 2016, 8] or climate control systems [Ramsey 2007, 136]): mental representations explain intelligent behavior by (and only by, it would seem) contributing within the framework of an architecture (broadly speaking), the functioning of which is governed by, or includes, certain basic operations – which might take a variety of forms, depending on the architecture in question, and tend, at least in the details of their operation, to be absent from other parts of the natural world.

Third, to a very great extent, the data of cognitive science are relational; they are outputs typed in terms of what they produce beyond the boundary of the organism: sound patterns

in the air, physical interactions *with objects*, navigation *through environments*, and reactions to *stimulus items on a monitor*. Various internal units contribute differentially to the production of such relationally typed behavior, where a unit's differential contribution correlates with certain external *relata*; and, against the backdrop of a successful theory of mental content, the activation of a given unit contributes significantly to the production of behavior that is directed toward (and thus partly individuated by its relation to) the very external object or structure that the unit in question represents. In which case, so long as the previously discussed conditions are met, the unit plays the role of a *mental* representation and contributes *as such* to the explanation of behavior – regardless of whether we can make sense of its contribution *qua* mere representation.

A simplistic illustration might help to clarify this aspect of the proposal (Rupert 2011). Assume that the cognitive system contains a unit the activity of which covaries with (in whatever way is posited by one's preferred causal-informational theory of content) the appearance of the subject's acquaintance John, and that moreover, this unit contributes significantly to the production of intelligent behavior directed at John (the subject's saying "Hi, John" upon John's entering the room, the subject's giving to John the same item that John asked to borrow on the occasion of a previous encounter, etc.). In this way, the unit contributes to the successful modeling of the subject's interactions with John and does so in virtue of its representation-related properties: of tracking, corresponding to, and directing behavior toward the thing tracked as the result the unit's participation in cognitive processing (that is, the kind of processing that distinctively accounts for the *explananda* of cognitive science). And, to the extent that a model running processes defined over symbols typed in accordance with what they refer to accounts for more variance in the data than competing models, the model's success helps to confirm the theory of content that assigns such referents – for instance, John, in our toy example.[12]

Misrepresentation can play an explanatory role as well. A successful model should account for the similarity of John-directed behavior to instances of behavior in which John does not appear as *relatum* (e.g. when the subject says "Hi, John" to Tom, a stranger). It will likely do so by exploiting the contribution of the unit that, across many contexts, differentially contributes to the production of John-directed behavior; the appearance of that John-representing unit also correlates with the presence of the sounds waves in question, as external *relata*, which helps to explain even data that do not directly involve John. Of course, one might think that modelers would get more explanatory mileage out of a theory of content that assigns "John or Tom" to the unit in question; but that would be to flout the well-advised proscription against statistical over-fitting; the proposed model would account for more of the variance in the limited data set (the subject's having said "Hi" to John on many occasions, and his now saying "Hi" to Tom), but would almost certainly come up short when put to predictive or explanatory use vis-à-vis new sets of data, as normally happens in cases of over-fitting.

Consider the contrast between the view developed above and Gładziejewski and Miłkowski's recent characterization of detector-based representations. They write, "Detectors are functionally bound to their targets … Detectors tend to react exclusively to certain states of affairs, and, in turn, generate cognitive or behavioral responses that are appropriate given the circumstances" (2017, 349) and "[T]he fact that S-representing, as a strategy of guiding action and cognition, is not purely reactive (as is the case with mere detectors), but involves an endogenous source of control" (2017, 339). Now, Dretske goes to some lengths, in the final chapters of *Explaining Behavior*, to show how indicator-based representations can contribute in complex ways, within the framework of the cognitive system, to the generation of a wide range of forms of behavior; so it does not seem fair to treat detector-based

representations as merely reactive, behavior-generating triggers, along the lines of the patellar reflex – at least not without showing where Dretske goes wrong. And, more to the present point, Gładziejewski and Miłkowski's description of the contribution of detector-based representations does not jibe with extant explanatory practice in cognitive science, which treats such detectors as units that partly constitute compound structures that are transformed by operations, series of the applications of which produce contextually appropriate responses. Take Ramsey's motivating example of detector-based representations in cognitive science, that of various feature-detecting cells identified by Hubel and Wiesel (Ramsey 2007, 119–120). The stimulation of these cells is, of course, selective to the stimulus features that cause their firing (orientation of a visually detected edge, in the case of one sort of cell of interest); but the contribution of such cells to the behavior that ensues is not functionally bound to those targets in the way that Gładziejewski and Miłkowski seem to be claiming. The activation of orientation-specific cells in V1 constitutes one step in a long process of information-extraction that builds, step-by-tedious-step, visual models (even if only partial, sketchy ones) that contribute to an exceptionally wide range of possible forms of behavior within the context of the cognitive architecture. It is only because the system can visually represent edges that it can visually represent tables and can then respond by doing anything from putting a coffee cup on the table to using the table to bar the door against intruders. The detector is not *solely responsible* for any of these relationally individuated forms of intelligent behavior, but it contributes to them all within the context of an architecture that functions according to distinctive principles; and it contributes in virtue of its representation-related capacity to track an oriented edge.[13]

Cognitive science addresses distinctive *explananda* – nonrandom scores on reading comprehension tests, the production of diagrams that eventuate in the appearance of buildings with the same structure as the diagrams, and so on. The mental representations in question contribute to accounts of those *explananda* only via their contribution within the context of a functioning architecture. Moreover, the architectures themselves function by principles that are not present throughout the natural world, even if the basic materials helping to determine the content of mental representations (being information-bearing, for instance) are ubiquitous; such principles range from "write and rewrite in binary code and transfer to a memory address" to "alter the connection strengths that contributed to this error" to "generate random vectors and interleave them with recently encoded vectors to prevent catastrophic interference."[14] Additionally, what the units contribute to these distinctive models is closely enough connected to our pretheoretical notion of a representation to be worthy of the label "mental *representation*"; for example, they contribute their capacity to track the very things they direct intelligent behavior towards (Rupert 2011) and they participate in processing that appears at least roughly inferential.

V. Engaging with the skeptics

Ramsey expresses the worry that detector- or indication-based representations function as mere causal mediators, intermediaries, or relays (Ramsey 2007, 125, 140, 149, 2016, 8); thus, he worries, even if they have some representation-related properties, they do not contribute such properties in the context of cognitive-scientific explanation. Here is Ramsey's argument, fleshed out: Causal-informational theories of representational content appeal to relations that appear throughout nature, far beyond contexts in which representational content plausibly appears (e.g. in the immune system or in physical mechanisms such as a thermostat); if a naturalistic theory assigns content only on the basis of such widespread relations, then the supposed content in question is not distinctively representational, and any

explanatory role apparently played by such content is not representational (Ramsey 2007, 142–145); if these representations *were* contributing *as such*, in virtue of their satisfying the job description of representations, then structures that clearly are not representations would also satisfy the job description of representations (which entails a contradiction); but any theory of representational content must deliver a form of representational content that plays a causal or explanatory role *as such;* thus, causal-informational theories fail. And, to be clear, the central charge of over-generalization (and thus failing to contribute in virtue of having met the proper job description of representations) is meant to apply even if a causal-informational theory is qualified, Dretske-style, so as to require that the structure with content have an acquired function (ibid. 131–132; Orlandi 2014, 117) or focuses on only on such structures as they appear in complex systems (Ramsey 2007, 145).

As powerful as such reasoning might seem at first blush, something *must* have gone wrong, for the form of reasoning applies to virtually all theoretical constructs beyond fundamental physics. If current physics is on the right track, the universe ultimately contains only a handful of forces and particle types. Thus, with regard to any domain *D* other than fundamental physics, the forces at work in driving any causal process in *D* will be forces that are also at work outside of *D* (because such forces are at work virtually everywhere). So, whatever the *D*, so long as it is not fundamental physics, Ramsey's reasoning categorizes its distinctive constructs as illegitimate, because they do not play a role distinctive of their job description. For example, electrostatic forces play a role in DNA transcription. Thus, whatever constructs are deployed in the study of DNA transcription, if the account of those constructs makes any substantive appeal to the contribution of electrostatic forces, those constructs are not really playing an explanatory role *as such* in molecular biology. After all, electrostatic forces permeate the universe, playing a role in contexts far beyond molecular biology. And, this pattern will appear through the natural world, regardless of the domain, so long as it is not fundamental physics. But, such a conclusion – that none of the constructs of sciences beyond fundamental physics play an explanatory role as such, and thus should be eliminated from our ontology – is patently absurd, at least from the standpoint of a naturalistic philosophy of science. At the very least, this is a huge bullet that neither Ramsey nor Hutto and Myin have shown any inclination to bite.

Thus, one cannot show that detection-based mental representations are illegitimate causal-explanatory constructs in cognitive science simply on the grounds that one of their central properties is determined partly by forces and relations – causation, information-carrying, nomic dependence – that appear outside of the representational context.

Ramsey (personal communication) hesitates to express his concern in the way I have immediately above. His fundamental worry is that the content-grounding materials appealed to by causal-informational theorists play the same role in cognitive processing – the role of mere causal mediation – that they play when found in patently nonrepresentational contexts. It is at this point where I think Ramsey's reasoning goes wrong. A mental representation with causal-informational content need not play such a minimal role. Being a detector-style mental representation requires much more than being a mere causal mediator. In fact, it requires much more than entering into such relations and having an acquired function. It requires that conditions (a), (b), and (c) be met, and mental representations with causal-informational content can (and, I think, most likely do, in the actual human case – though establishing this is not my purpose here) meet these three conditions. Moreover, causal mediators appearing beyond the cognitive context do not meet these conditions, because they do not contribute to the production of intelligent behavior by their participation in operations appealed to distinctively in cognitive modeling.

Returning to the general case, when faced with causal-explanatory constructs outside of fundamental physics, we should ask such questions as "What is the overall structure of the units, entities, or frameworks appealed to in the distinctive causal-explanatory enterprise in question?" "What distinctive *explananda* does this structure account for?" and "What distinctive processes govern the formation or deployment of the component units in question, as they appear in this structure, so as to help to account for the *explananda* distinctive of the enterprise in question?" The answers to these three questions establish that, say, *being a nucleic acid* is not an illegitimate construct – failing to contribute *as such* – regardless of the fact that the bonding patterns of nucleic acids are governed partly by electrostatic forces, which also happen to play a role outside of the context of DNA transcription.

What about the case of mental representations? The Dretskean responses to these three questions would seem to be that the units in question help to account for action (the distinctive *explananda*) by being part of a larger integrated system in which motor control processes, beliefs, and desires appear (i.e. the further structure distinctive of the domain), the relevant relations between the parts of which are governed by reinforcement learning (the distinctive process). For my part, I would translate this into the cognitive-scientific context, arguing that the distinctive *explananda* comprise a wide range of measurable forms of intelligent behavior of the sort mentioned above (involving everything from holding a conversation, to planning and constructing a building, to devising an experimental apparatus), that the added structure is a cognitive architecture (broadly understood), and that the distinctive processes are those governing the relevant operations in the architecture (such as backpropagation or storing in binary code at an address in a memory structure – and very many other candidates appear in cognitive-scientific modeling). Given that the present debate concerns the use of what strikes many as a pretheoretical term, it might also be required that the way in which the units in question play the role just outlined reflect well enough the pretheoretical conception of representation: a given unit corresponds to the thing it represent; it does so in a way that admits of misrepresentation (or inaccuracy); it distinctively governs actions immediately directed toward the thing represented (by tracking it, for example) and distinctively governs actions *in*directly aimed toward the thing represented (by being used as a stand-in for the thing represented during so-called off-line cognition – in the case of detectors in the visual system, for example, some are used in the exercise of imagination); and its participation in processing distinctive of cognition is at least roughly inference-like.

Ramsey and Hutto and Myin seem to miss this way of looking at things because they fail to emphasize the central components of Dretske's framework: reinforcement learning, within the life of the organism, and its effect on the behavior of the organism that such reinforcement learning is meant to explain.[15] And setting aside questions about their exegesis of Dretske's work, when one transports these Dretskean ideas into the realm of cognitive-scientific modeling, the "mere causal mediator" problem, and the associated "as such" problem, disappears.

An appreciation of the way in which the central components of Dretske's theory interact should also stifle a further criticism of Dretske's account expressed by both Ramsey and Hutto and Myin, that there's something inherently confused about the notion of information or indication as the foundation of a naturalistic theory of mental content. There's nothing confused here. One state's carrying the information that another state obtains is defined precisely, as reduction of uncertainty at the source, which is itself expressed mathematically (Dretske 1981, chapters 1–3).[16] Ramsey (2017, 4202) complains that information is not some sort of substance. Fair enough, but it is not clear which information theorist claims otherwise. The Weaver-Shannon equations governing the quantities of information

transmitted are not offered as an account of a flowing substance, such as one might find in a textbook on fluid dynamics.[17] And, when the time comes to specify with any precision what information processing amounts to, of the sort that might be useful in cognitive science, the metaphor of a flowing substance seems nowhere to be found. Russell and Norvig (2009) write over a thousand pages about models in artificial intelligence, carefully detailing various forms of information processing; but none of what they write constitutes, or is intended to constitute, a theory of a substance that flows.

Both Ramsey and Hutto and Myin make great critical hay of Dretske's use of such terms as "say" and "tell," resting objections to causal-informational theories on intuitions associated with everyday uses of these terms (Hutto and Myin 2012, 62, 65, 67; Ramsey 2007, 135, 136, 138, 139). But, nothing in Dretske's framework requires any bit of the cognitive system literally to say or tell anything to anyone. Dretske puts both terms, "say" and "tell," in scare quotes (1988, 98–99 and 90, respectively). Dretske is here using informal, accessible language to try to get across a "big idea," but the gist of his theory – the bits that are laid out in precise detail and the bits that should be subject to critical evaluation – appear in Dretske's more careful, mathematical presentation.

An emphasis on technical details can help us to identify the flaw in a related criticism of Dretske's view. The dialectical backdrop is somewhat complex, but here is the gist of it: Ramsey asks whether information is or is not distinct from nomic dependence; then, by dilemma, he argues that the as-such problem arises. Ramsey argues by example, one he takes to have application on either horn of the dilemma. (My objection to Ramsey's use of the example applies regardless of the horn; thus, in what follows, I do not attend much to the dilemma structure.) A subject plants a tree in his yard in order to provide shade. That, Ramsey says, is a way of exploiting nomic dependency – of the sort that permeates nature – and thus surely does not deliver representational content that plays a role; speaking as the subject, Ramsey says of the tree, "[I]t doesn't tell me anything" (2007, 136–138). But, let's apply Dretske's actual framework, setting aside the informal use of "tell." A given human subject has, we can assume, what we might naturally call a "coolness" detector. Now, given what we know about thermal receptors (Akins 1996, 346 ff), what we might think of as the peripheral detector of temperature actually carries information about a wildly gerrymandered range of states. Perhaps, the detector of actual interest appears downstream, then, such that the environmental event it indicates is less gerrymandered, that is, contains fewer distinct outcomes in the information space, although one of those outcomes is relative coolness in the environment. No matter. Wherever we look in this stream of processing, we are likely to find an indicator that carries the information "cool temperature *or* … *or* … " Let us say now that, on a particular occasion, the environmental state co-present with the activation of the indicator in question is, in fact, a cool temperature, and the subject responds by moving under a tree; and as a result of moving into the tree's shade, the connection between the indicator in question and movement directed toward shady areas is strengthened. On Dretske's view, that indicator becomes a representation of cool temperature (perhaps only after number of successful movings caused by the activation of the detector in question – depending on, among other things, the value of the learning-rate parameter in the relevant reinforcement equation). Through the chaining of such representational structures (see Dretske 1988, chapter 6, for his take on the details), the content of this representation can contribute to the production of intelligent behavior, the later planting of a new tree, which will then provide shade to be enjoyed by the subject, even though the indicator in question indicates a collection of possible states of the world that includes states other than relative coolness.[18]

VI. Diagnosis

This section offers a partial diagnosis of the source of disagreement. In broad strokes, it results from critics' apparent inclination toward a robust description theory of reference for scientific terms. To be sure, descriptions play a role in determining the reference of scientific terms, but such descriptions tend to be relatively vague and shifting, oriented toward success in accounting for the phenomena to be explained, rather than reflecting a substantive pretheoretical understanding of the properties, states, or kinds in question (Wilson 2008). In contrast, the skeptics seem to have in mind too definite a set of descriptions, tied too tightly to a folk-psychological understanding of the mind and mental processing.[19]

This commitment manifests itself partly in the demand for a job description for mental representations as a substantive and binding constraint on what can count as representations, independent of the observable, measurable *explananda*. The as-such problem grows out of this challenge. Once the job description of representations has been sufficiently clarified, the thought goes, we can tell whether an entity or structure plays that role and contributes to explanation in virtue of playing that role.

Now, contrary to what Ramsey claims (e.g. 2016, 5, 7), leading naturalists have always been sensitive to the need for a job description for mental representations. Those who address this question, in some significant way, amount to a veritable who's who among naturalistically oriented philosophers of mind working in the 1980s and 1990s: Dretske (1988, chapter 3), Millikan (1984, chapter 6), Cummins (1996), Clark and Toribio (1994), Haugeland (1991), and arguably Fodor (1987, chapter 1).[20] Ramsey himself acknowledges many of these counterexamples to his historical claim. But, as the counterexamples stack up, Ramsey, somewhat inexplicably, remains unmoved by that fact (2016, 5, 7).

My guess is that what is bothering Ramsey is not that prominent naturalistic philosophers of mental content have not attended to the job description challenge. Rather, it is that many of them have had too vague a description in mind (by Ramsey's lights). Scientific theory construction is generally not constrained by a description theory of reference, such that one's scientific proposal about the properties of *x*'s cannot *really* be a theory of *x*'s unless the entities proposed meet all of the necessary conditions that appear in a detailed, pretheoretically formulated description of *x*, at least when "x" is a term with pretheoretical use (cf. Dretske's comments about a scientific theory of information – 1981, 46–47). Hutto and Myin and Ramsey seem to have lost sight of this – either in their demand that there be an antecedent, constraining description or in the inclusion of, and attachment to, specific conditions in such a description. They themselves might be committed to a sort of conceptual analysis that presupposes a description theory of reference, but, be that as it may, they should not use such a commitment as leverage against the naturalist.[21] Thus, perhaps it is not that naturalistically oriented philosophers of mental representation have failed to address the job description challenge; perhaps it is that, in many cases, they have addressed it only in the relatively minimal way – in fact, in a way appropriate to the naturalistic treatment of terms referring to theoretical kinds, properties, etc., rather than in the way that Ramsey and Hutto and Myin think a naturalist should.

Consider this point in connection with Hutto and Myin's concern about intensionality (with an "s"). A long-standing philosophical intuition – extant at least since Frege's "On Sense and Reference" – holds that representations have a sense, mode of presentation, or some other semantic property that can come apart from the referent or what is represented; in the extreme case (in one's mental representation of Santa Claus, for example), this extra-referential component of meaning appears in the complete absence of a thing referred to. On Hutto and Myin's view, *being intensional* (with an "s") is essential to content; any

theory of content, naturalistic or otherwise, that does not posit an intensionality-supporting semantic property fails (2012, 79). But, so far as I can tell, this reflects a misunderstanding of how descriptions constrain the reference of scientific terms.[22]

Consider the following example of intensionality-related behavior. Take a hypothetical subject Alex, who asserts all of the following sentences:

(1) Hesperus appears in the evening.
(2) Hesperus is not identical to Phosphorus.
(3) Phosphorus appears in the morning.
(4) Phosphorus, low in the morning sky, is a sight to behold.

An interlocutor asks Alex, "Would you like to go out with me at 5:00 tomorrow morning to view Hesperus?" In response, Alex rolls his eyes and exclaims "Hesperus isn't out in the morning! But, I'd love to go see Phosphorous with you tomorrow morning." The externalist contents of the beliefs expressed by (1) and (3) are identical by hypothesis (that is, (2) is false). So, why would Alex express conflicting attitudes toward one and the same object?

Here is one naturalistic account (adapted from Fodor 1990; though, for the roots of the idea, see Fodor 1981, 188, 194; see also Rupert 2008b, 2013).[23] The syntactically individuated mental representations that produce verbal outputs "Hesperus" and "Phosphorous," respectively, differ in the various causal and associative relations they enter into with respect to other syntactically individuated mental representations, in particular, in their relations to the syntactically individuated strings that produce the verbal output "is visible in the morning" and "is visible in the evening." Alex's knowledge structure associated with the syntactically individuated unit that produces utterances of "Hesperus" does not include the syntactically individuated string that produces utterances of "appears in the morning," while the knowledge structure associated with Alex's syntactically individuated unit that produces utterances of "Phosphorous" does. These aspects of the model of Alex's cognitive system, along with a host of other aspects of the model (e.g. assumptions about processing steps in his cognitive architecture) account for the data – that is, the intensionality-related verbal behavior – in question.

The preceding might not reflect one's pretheoretical expectations concerning the details of an explanation of intensional behavioral phenomena; one might expect an explanation that appeals more directly to the content of the representations involved. But, why is that expectation relevant to the enterprise of cognitive science? We have a guess about what sorts of things a completed cognitive science should account for. Among them are odd variations in pairings of linguistic utterances and contexts. And we might guess that our best models will include sense-like contents as primary determinants of that intension-related behavior.[24] If Fodor is right, though, things turn out otherwise; instead, the behavior is driven by the computational–syntactic properties of vehicles, and the way the architecture associates these vehicles with each other, given the subject's history. From the standpoint of philosophy of science, that kind of result is unsurprising. It should not tempt us to say, "Aha, the units in question are not mental representations, after all." Their status as mental representations depends on the role they play in cognitive modeling more generally, per (a), (b), and (c), above.

I do not claim to have in hand a precise account of the conditions under which an entity posited by some successful science and referred to as "x" has become too far detached from our pretheoretical conception of x's as definitely not to count as an x. The continued usage of a term in the sciences – whether a term borrowed from everyday language or a technical term – often follows a messy, wandering path (Wilson 2008), and the history of successful

science reveals no algorithm that determines when to preserve a term with shifting significance and when to eliminate its use, or determines when to continue to use the term while acknowledging that it refers to something other than what it had previously referred to.

It is important to keep in mind, however, the *explananda*'s central role in the process; often, relative continuity in the *explananda* accounted for helps to drive the decision to retain a given term, even in the absence of relative continuity in the properties associated with what we take to be designated by that term. Cast in the mold of a description theory, we might recommend the following descriptive guidance: mental representations are things that help to account for the intelligent forms of behavior by exhibiting some significant number of the properties associated pretheoretically with representations (such properties as standing in for, tracking, or corresponding to in a privileged way, having the capacity to misrepresent – themselves properties our conceptions of which might change over time) and by participating in processes (such as backpropagation of error) that are distinctive of the overall structures that explain intelligent behavior. This might not be *the* correct, reference-guiding description for "mental representation"; but I would maintain that it captures any reasonable demand on mental representations that play a role *as such*. And, detector-based mental representations certainly can fit this description.

In closing, I say a bit about what I take myself not to have done. I have not attempted to show that, in addition to structures that merely enter into relations of nomic dependence or information-carrying, on the one hand, and mental representations, on the other, that there is a third category, consisting of representations neat. Perhaps, the only form of representation in our universe is mental representation. Or, perhaps representations appear far and wide – either as an expansive proper subset of the natural relations listed above (far more expansive than set of mental representations) or because being a representation is coextensive with being one of the *relata* of an instance of one of those natural relations, in which case representations permeate the universe. I have not taken a stand on such matters, and for the purposes of this paper – to establish the viability of a causal-informational theory of the content of mental representations in the face of the job description and as-such challenges – taking a stand on such questions is unnecessary.

Moreover, I have not tried to show that mental representations, in fact, appear in human minds or cognitive systems (see note 1). I would bet that human cognition is representational, partly because I think the proposal I have advanced jibes well with our best current cognitive science. But, things might turn out differently; our best models of the production of intelligent behavior might, for instance, represent such behavior as the result of forms of dynamical processing not at all distinctive of the cognitive domain. In which case, it would seem to follow from my proposal that human minds do not traffic in mental representations. This question remains open, for all that I have argued here.

Acknowledgements

Many thanks to Bill Ramsey, Heather Demarest, Bob Pasnau, Raul Saucedo, and an anonymous reviewer for comments on earlier drafts of this paper; to Frances Egan, Tobias Schlicht, Krzysztof Dołęga, and Paul Schweizer for discussion of the issues addressed herein; and to an audience at the University of Edinburgh for helpful feedback.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes

1. And, to be clear, I do not here argue that detector-style representations play a definite causal-explanatory role in cognitive science. My brief is more modest, partly because space is limited. My goal is to neutralize the sort of in-principle roadblocks to causal-informational theories of mental content that Hutto and Myin and Ramsey take themselves to have identified. I do so partly by outlining, at a fairly abstract level, a causal-informational approach not subject to Hutto and Myin and Ramsey's concerns. But, there may well be other roadblocks to the development of a causal-informational theory of mental content; that is, there may be distinct reasons to think that naturalized, detector-like representations (those subject to a causal-informational theory and lacking internal cognitive structure) cannot or will not play a causal role in a science of human cognition. I take no stand on such matters here. Partly for this reason, I make only passing reference to various aspects of cognitive-scientific modeling and do not focus on any particular research program or family of research programs to try to show that representations do, in fact, play a causal-explanatory role in cognitive-scientific modeling. My strategy takes, as a starting point, the rampant use of "representation," especially in cognitive neuroscience, to refer to cognitively unstructured detectors (bits of cortex that "light up" in response to certain sorts of stimuli, for instance); I then ask whether Hutto and Myin's and Ramsey's criticisms show that such uses of "representation" are inherently confused or mistaken.
2. In much of Fodor's writing (e.g., 1975, 1981, chapter 10, 1983, 1998), he engages directly with empirical results in cognitive science (in fact, he has himself published experimental work – Fodor et al. 1980) and trends in those results that are of theoretical import from the standpoint of cognitive science itself. So, in one respect, what is written in the main text might seem to misrepresent Fodor's work vis-à-vis cognitive science. Such a reaction would, however, reflect a misunderstanding of Fodor's work on mental content (1987, 1990), that is, his work as what I below call a "first-generation" naturalistic theorist of content. In that work, his primary goal is to provide a naturalistic semantics for folk-psychological states that is respectable from the standpoint of cognitive science. More generally, in much of Fodor's work, his explicitly stated goal is to find, in cognitive science, a vindication of folk psychology (he uses the language of "vindication" throughout chapter 1 of *Psychosemantics*), a situation that will help to confirm the very science in question, given a wealth of independent reasons (both empirical and, loosely speaking, *a priori*) for accepting folk psychology. The structure of this aspect of Fodor's thinking emerges at least as early as "Propositional Attitudes" (Fodor 1978):

> "What are the propositional attitudes?" … One way to elucidate this situation is to examine theories that cognitive psychologists endorse, with an eye to explicating the account of propositional attitudes that the theories presuppose. That was my strategy in Fodor (1975). In this paper, however, I'll take another tack. I want to outline a number of a priori conditions, which, on my view, a theory of propositional attitudes (PA) ought to meet. I'll argue that, considered together, these conditions pretty clearly demand a treatment of PAs as relations between organisms and internal representations; precisely the view that the psychologists have independently arrived at. I'll thus be arguing that we have good reason to endorse the psychologists' theory even aside from the empirical exigencies that drove them to it. I take it this convergence between what's plausible a priori and what's demanded ex post facto is itself a reason for believing that the theory is probably true. (501)

More to the present point, Fodor describes the goal of his first major work on content, *Psychosemantics*, in the following way: "This book is mostly a defense of belief/desire psychology" (Fodor 1987, *xii*). And, at greater length:

> The main thesis of this book can now be put as follows: We have no reason to doubt – indeed, we have substantial reason to believe – that it is possible to have a scientific psychology that vindicates commonsense belief/desire explanation [i.e., folk psychology]. But though that is my thesis, I don't propose to argue the case in quite so abstract a form. For there is already in the field a (more or less) empirical theory that is, in my view, reasonably construed as ontologically committed to the attitudes and that –

> again, in my view – is quite probably approximately true. If I'm right about this theory, it is a vindication of the attitudes. (Fodor 1987, 16)

In fact, Fodor spends the first ten pages of *Psychosemantics* defending the depth, the success, and what he claims is the indispensability of folk psychology (appealing, *inter alia*, to our ability to understand the plot twists in Shakespeare's plays and the inferences drawn by Sherlock Holmes, as well as our patent ability to coordinate everyday social interactions). It should come as no surprise that Fodor's development of a theory of content was driven largely by the quest to vindicate folk psychology. One of the most lively and high-profile debates in philosophy of mind in the 1980s, which Fodor was at the center of, concerned the questions (a) whether folk psychology is legitimate, (b) whether its being legitimate requires that it mesh with, or be vindicated by, cognitive science, (c) what kind of cognitive science, if any, would be required to vindicate it, and (d) whether that kind of cognitive science is in the offing (Churchland 1981; Dennett 1987; Fodor 1987; Stich 1983).To muddy the waters a bit, Fodor recognizes that our best cognitive-scientific models include subpersonal-level representations distinct from, and in addition to, those that play a direct role in vindicating folk psychology (that is, distinct from the computational symbols that encode propositions that are the contents of beliefs and desires and that play the distinctive causal roles of beliefs and desires) (1987, 25–26); nevertheless, he does not develop a theory of content specially aimed at such additional representations, presumably because their semantics is tangential to his project in *Psychosemantics*, of vindicating commonsense belief–desire psychology.

3. To be clear, naturalists can stipulate just about *anything* to be a potentially interesting relation, and then let the stipulation run the empirical gauntlet; but they cannot simply stipulate just anything to be mental content or to be a mental representation. There has to be *some* principled connection to the pretheoretical use of "content" or "representation," more on which below.
4. Contrast this approach with the shape of Fodor's approach in the section "The Essence of the Attitudes" (1987, 10–16), which is to identify, *a priori*, the essential properties of the folk-psychological constructs of belief and desire and then to look for scientifically respectable states having those properties.
5. The comments in the main text oversimplify the history in some respects. All of the first-generation authors address or incorporate, in some way, work in cognitive science (see note 2's relevant remarks about Fodor's particularly complicated situation in this regard); but they tend to focus on a cognitive science the purpose of which is to illuminate or ground folk categories, for example, "thought and experience" (Dretske 1997, 4). In contrast, second-generation theorists are, to a significantly greater extent, concerned to a develop a naturalistic theory of content that suits the needs of working cognitive science, as the endeavor to model the mechanisms that produce the relevant measurable, replicable data – leaving the status of folk psychology to fall where it may. But, they do not focus exclusively on this; Ryder, for example, would like his neurosemantic theory to account directly for *explananda* of both scientific and folk psychology (2004, 212, 232).
6. Ramsey sometimes seems to endorse the naturalistic perspective associated in the main text with Cummins's work and the work of second-generation naturalistic theorists of content (Ramsey 2007, 65–66). Nevertheless, throughout his discussion of Dretske's approach, Ramsey repeatedly makes a critical appeal to such claims as that information is supposed to "tell" (2007, 135, 136, 138, 139) someone something or "inform" (ibid., 141, 148) an agent of something, which flies in the face of the methodological point in the main text. Dretske gets to stipulate what the information-relation is, as he does in painstaking detail (1981, chapters 1–3); when the relation, so stipulated, holds, the relevant relatum carries such-and-such information. Of course, the question might then be asked whether that relation, as stipulated, does the explanatory work Dretske wants it to do, as the naturalistic basis for, for example, the everyday use of belief talk to explain behavior.
7. Compare Matthen's claim that "[C]ontent is a system property. It is a property of states in a system that treats information-carrying states in a certain way" (2014, 125). Also see Markman and Dietrich (2000, 144), for the claim that Dretskean information-bearing mediating states become representations only when they play a specific kind of role in a cognitive system (allowing the system to satisfy its goals, for example).
8. To be clear, then, the purpose of the preceding section was not to try to show that Dretske gets things right. Rather it is to show that Ramsey's and Hutto and Myin's criticisms of Dretske's

view fail to engage fully with it and that a careful look at the structure of Dretske's view orients naturalistic theorists of mental content in the right direction, that is, it reveals what a more promising naturalistic semantics for detector-style representations might look like. In other words, although I do not think Dretske has the details right, and am thus not out to defend the details of his proposal, his theory instantiates a structure that others working on naturalistic theories of content can fruitfully extract and flesh out differently.

9. Dretske distances himself from cognitive-scientific explanation, holding that the value of his approach lies in the domain of folk psychology (Dretske 1988, 81 n1).
10. On the importance of the *explanandum*, see Dretske 1988, 69; on the distinction between representation and mental representation, see Dretske 1997, 19.
11. My particular interest is in the vindication of causal-informational theories as they might apply to detectors, indicators, or otherwise unstructured cognitive units. This is not meant to marginalize what Ramsey calls "S-representations" (2007, chapter 3), that is, complex structures with map-like properties or, more generally, structure that mirrors or simulates the structure of the space of the real-world problem to be solved (cf. Cummins 1996). There is ample evidence of the brain's use of such structures. See Ramsey (2016) for some discussion of how S-representation and the detector-based notion of representation might be wedded productively.
12. In this vicinity, one may find a realist response to Frances Egan's skepticism about mental representation, expressed within the framework of computational cognitive science (Egan 2014; also see Schweizer 2017). Egan claims that computational cognitive science is nonrepresentational partly because, I think, she ignores the relational nature of the data and focuses on the mathematical theory of computation. But, cognitive science is, in some clear sense, an applied science – a science of relationally individuated data. In this applied context, the characterization of computational units as representations is no more a mere gloss than is the relational characterization of the data, that is, not a mere gloss at all.

    So far as I can tell, Egan's tendency to view the data as nonrelational results partly from a mistaken view about the role of modal commitments in natural science. It is not the case that naturalistic explanation requires that, in order for (process, phenomenon, property, state) X to explain (process, phenomenon, property, state) Y, there is no possible world in which the apparent relations between X and Y are, relative to the actual world, shuffled around. There is no evidence that scientific explanation is driven (with regard to evidence and justification, as opposed to brainstorming) by such modal concerns. In which case, the sorts of examples Egan uses to motivate a merely-gloss-based approach to representation – the Visua example, for instance (ibid., 126–127) – seem beside the point, for they rest on claims about which internal states (processes, etc.) could be paired with which external states (processes, etc.) in distant possible worlds.

    The application of the present line of thinking to Egan's work demands much more attention than this, but given limitations of space and the extent to which Egan's concerns are removed from those of prominent enactivists, more extensive discussion must be postponed.
13. It is largely because the theory of mental representation builds content from representation-related, representation-like, or vaguely representational ingredients that it is not infelicitous to use the term "mental *representation*." But, the use of that term in no way implies that wherever one finds representation-like ingredients, one has found a representation *simpliciter*. It might be that there is no unified genus *representation*. Perhaps instead there are only some basic natural ingredients that loosely fit some of our vague intuitions about representation, which ingredients then, when put to use in a particular applied science *x*, become *x* representations. On this view, intuitions concerning what counts as playing a representational role might provide sufficient unity to the use of "representation" across these various specific contexts, without there being any naturalistically relevant kind *representation neat* that plays a causal-explanatory role in any science at all. Thus, one can be in a position to offer a compelling naturalistic account of mental representation without committing oneself to there being a definite, useful, or genuine kind, *representation*.
14. One might reinterpret Dretske's folk-psychology-oriented strategy in this light, by treating reinforcement learning – with its array of motivational states and strengthened associations between indicators and motor commands – as the distinctive operation.
15. Compare the way in which Ramsey runs his head-to-head comparison of detectors and icon-like S-representations (2007, 194 ff.). Although Ramsey considers the possibility that the detectors in question acquire a function, he makes little mention of the sort of reinforcement learning or

the appearance of structuring causes of importance to Dretske. Moreover, Ramsey's central example – a car with sensors at its periphery – is not a cognitive system, that is, one that displays a wide variety of forms of intelligent behavior. Thus, Ramsey's head-to-head comparison has no clear bearing on matters at hand, that is, on questions about *mental* (or cognitive) representation.

16. In *Explaining Behavior* (58–59), Dretske states that he does not want to presuppose the detailed account of information spelled out in from *Knowledge and the Flow of Information* (1981). But, he explicitly equates information and indication; what technical bits he does say about indication in (1988) reflect what he says in (1981) about information; and thus, there is no reason to think he has rejected the more detailed analysis. To the contrary.
17. After discussing some intuitively natural ways to think about the "information channel," Dretske writes,

    > From a theoretical point of view, however, the communication channel may be thought of as simply the set of [statistically defined] dependency relations between *s* and *r*. If the statistical relations defining equivocation and noise between *s* and *r* are appropriate, then there is a channel between these two points, and information passes between them, even if there is no direct physical link joining *s* with *r*. (1981, 38)

18. Here is a bit more about Ramsey's disjunctive premise – that information *is* nomic dependence or is wholly distinct from nomic dependence – which provides the backdrop for his dilemma argument. According to Dretske, the carrying of information amounts to the obtaining of the conditional probabilities in question relative to the state of the receiver (1981, 68), so long as those probabilities holds in virtue of natural law (1981, 76–77). Thus, even if the pattern of conditional probabilities central to the definition of information holds, if it does not hold because of nomic regularities, then the receiver states in question do not carry the information they would otherwise. Thus, it is neither the case that information *is* nomic dependence nor is wholly distinct from nomic dependence; rather, the correlations constitutive of information are "a symptom of lawful connections … information inherits its intentional feature from the nomic regularities on which it depends" (ibid., 77). It is the obtaining of a pattern of conditional probabilities because of nomic regularity. And, that is precisely what is exploited in the use of shade, within the broader structural and temporal context that brings a mental representation into existence.
19. Ramsey sometimes seems to take an appropriately flexible, pragmatist approach to the understanding of scientific processes. For instance, he views the nature of cognitive science as something that – beyond a general conception of what cognitive science is out to explain – should be left largely open, to be negotiated as the project proceeds (2017, 4208–4209). In my view, that approach applies equally to the scientific conception of representation, as something that should not be rigidly characterized in advance, beyond perhaps some very general platitudes largely to do with what is to be explained.
20. Outside of philosophy, Markman and Dietrich (2000) offer a sophisticated take on the job description of mental representations (as part of a discussion that is strongly informed by, and that significantly contributes to, the project of naturalizing mental representation as a theoretical posit in cognitive science).
21. This is not to say that descriptions never drive the evolution of the use of scientific terms; see, e.g., Ian Hacking's (1983, 87–90) account of how the reference of "meson" was fixed.
22. Although I have explicitly identified myself as a second-generation naturalistic theorist of mental content, I might seem now to have "slid back" into the game played by first-generation theorists. The dialectic is a bit different, though. I certainly do think our naturalistic theory of content should be responsive, in the first instance, to the needs of causal-explanatory modeling of measurable data that produces replicable results, letting folk psychology fall where it may. Nevertheless, once one has identified such units and assigned to them privileged relations to other individuals, properties, or kinds (a relation one takes to be content-fixing), the further questions can be asked, "Are those units really representations?" and "Do they play a representational role?" At that point, even the second-generation naturalistic theorist of mental content must engage with the question whether those units satisfy enough of the intuitions associated with the everyday term "representation" as to be rightly called "mental representations," so as to avoid misleading stipulation. At that point in the debate, even the second-generation theorist must come to grips with at least some of the everyday intuitions about the use or application of "representation" and show that the cognitive-scientific models in question do

enough of the right sort of causal-explanatory work to count as representations. That is the task at hand in the main text; the reasoning of this section does not assume that cognitive science's goal is to vindicate folk psychology.

23. On this topic, Hutto and Satne (2015, 523) may slightly misrepresent the history. On their story, Fodor rejected senses only late in his career. But as early as 1990, Fodor wrote "The older I get, the more I am inclined to think that there is nothing at all to meaning except denotation … ." (1990, 161).
24. The problem of intensionality is often associated with the human ability to think about nonexistent things or kinds. One can think about unicorns, even though unicorns do not exist, and thus, it is often thought, there must be a component of meaning that is neither syntactic nor referential; this is commonly taken to be a sense or an intension. One naturalistic strategy is to treat such concepts as composed: when one thinks of unicorns, one is thinking of horses with horns, at a first approximation (Fodor 1990). This strategy takes on additional plausibility in the current context, in which emphasis is placed on a cognitive-scientific notion of representational content, the primary role of which is to account for largely relational data. After all, since there are no unicorns, there will be no data concerning interactions with unicorns for cognitive science to puzzle over. What is left is the production of "unicorn"-related sentences (reports of intuitions about unicorns, etc.), which, though relational in a sense (they involve the production of such things as sound waves beyond the boundary of the body), appear to be a much more manageable target for the proposed approach to intensionality described in the text – combined perhaps with the appeal to compositionality.

Notes on contributor

Robert D. Rupert writes about issues in philosophy of mind, the philosophical foundations of cognitive science, and related areas of metaphysics, philosophy of science, and epistemology. His book, *Cognitive Systems and the Extended Mind*, was published by Oxford University Press, and his work has appeared in *Journal of Philosophy*, *Noûs*, *Mind*, *Philosophy and Phenomenological Research*, *Mind & Language*, *British Journal for the Philosophy of Science*, *Philosophical Studies*, *Synthese*, and many other journals and volumes. He has held visiting positions at University of Edinburgh, the Australian National University, and Ruhr University, Bochum, among other institutions. He is currently an Associate Editor at the *British Journal for the Philosophy of Science*.

References

Akins, Kathleen. 1996. "Of Sensory Systems and the 'Aboutness' of Mental States." *Journal of Philosophy* 93 (7): 337–372.
Chemero, Anthony. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.
Churchland, Paul M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78: 67–90.
Clark, Andy, and Josefa Toribio. 1994. "Doing without Representing?" *Synthese* 101: 401–431.
Cummins, Robert. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.
Cummins, Robert. 1996. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.
Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Dretske, Fred. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.
Dretske, Fred. 1997. *Naturalizing the Mind*. Cambridge, MA: MIT Press.
Egan, Frances. 2014. "How to Think about Mental Content." *Philosophical Studies* 170: 115–135.
Eliasmith, Chris. 2003. "Moving Beyond Metaphors: Understanding the Mind for What It Is." *Journal of Philosophy* 100 (10): 493–520.
Fodor, Jerry A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.
Fodor, Jerry A. 1978. "Propositional Attitudes." *The Monist* 61 (4): 501–523.
Fodor, Jerry A. 1981. *Representations*. Cambridge, MA: MIT Press.
Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.
Fodor, J. A., M. F. Garrett, E. C. T. Walker, and C. H. Parkes. 1980. "Against Definitions." *Cognition* 8: 263–367.
Gładziejewski, Pawel, and Marcin Miłkowski. 2017. "Structural Representations: Causally Relevant and Different from Detectors." *Biology and Philosophy* 32: 337–355.
Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
Haugeland, John. 1991. "Representational Genera." In *Philosophy and Connectionist Theory*, edited by William Ramsey, Stephen Stich, and David Rumelhart, 61–90. Hillsdale, NJ: Lawrence Erlbaum Associates.
Hutto, Daniel D., and Erik Myin. 2012. *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA: MIT Press.
Hutto, Daniel D., and Glenda Satne. 2015. "The Natural Origins of Content." *Philosophia* 43: 521–536.
Markman, Arthur B., and Eric Dietrich. 2000. "In Defense of Representation." *Cognitive Psychology* 40: 138–171.
Matthen, Mohan. 2014. "Debunking Enactivism: A Critical Notice of Hutto and Myin's *Radicalizing Enactivism*." *Canadian Journal of Philosophy* 44 (1): 118–128.
Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Orlandi, Nico. 2014. *The Innocent Eye: Why Vision Is Not a Cognitive Process*. Oxford: Oxford University Press.
Papineau, David. 1984. "Representation and Explanation." *Philosophy of Science* 51: 550–572.
Prinz, Jesse J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.
Ramsey, William M. 2016. "Untangling Two Questions about Mental Representation." *New Ideas in Psychology* 40: 3–12.
Ramsey, William M. 2017. "Must Cognition Be Representational?" *Synthese* 194: 4197–4214.
Rosenbloom, Paul S., John E. Laird, Allen Newell, and Robert McCarl. 1991. "A Preliminary Analysis of the Soar Architecture as a Basis for General Intelligence." *Artificial Intelligence* 47: 289–325.
Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (9): 533–536.
Rupert, Robert D. 1998. "On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought." *Synthese* 117: 95–131.
Rupert, Robert D. 1999. "The Best Test Theory of Extension: First Principle(s)." *Mind and Language* 14: 321–355.
Rupert, Robert D. 2004. "Challenges to the Hypothesis of Extended Cognition." *Journal of Philosophy* 101: 389–428.
Rupert, Robert D. 2008a. "Causal Theories of Mental Content." *Philosophy Compass* 3: 353–380.
Rupert, Robert D. 2008b. "Frege's Puzzle and Frege Cases: Defending a Quasi-Syntactic Solution." *Cognitive Systems Research* 9: 76–91.
Rupert, Robert D. 2009. *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.
Rupert, Robert D. 2010. "Extended Cognition and the Priority of Cognitive Systems." *Cognitive Systems Research* 11: 343–356.
Rupert, Robert D. 2011. "Embodiment, Consciousness, and the Massively Representational Mind." *Philosophical Topics* 39: 99–120.
Rupert, Robert D. 2013. "On the Sufficiency of Objective Representation." In *Current Controversies in Philosophy of Mind*, edited by Uriah Kriegel, 180–196. New York: Routledge.
Russell, Stuart, and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Pearson.
Ryder, Dan. 2004. "SINBAD Neurosemantics: A Theory of Mental Representation." *Mind and Language* 19: 211–240.
Schweizer, Paul. 2017. "Cognitive Computation sans Representation." In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*, edited by Thomas M. Powers, 65–84. New York: Springer. doi:10.1007/978-3-319-61043-6.

Shea, Nicholas. 2007. "Consumers Need Information: Supplementing Teleosemantics with an Input Condition." *Philosophy and Phenomenological Research* 75 (2): 404–435.
Shea, Nicholas. 2014. "Reward Prediction Error Signals Are Meta-Representational." *Noûs* 48 (2): 314–341.
Slater, C. 1994. "Discrimination Without Indication: Why Dretske Can't Lean on Learning." *Mind and Language* 9: 163–180.
Stewart, John, Olivier Gapenne, and Ezequiel A. Di Paolo, eds. 2010. *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press.
Stich, Stephen P. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Stich, Stephen P. 1996. *Deconstructing the Mind*. New York: Oxford University Press.
Usher, Marius. 2001. "A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation." *Mind and Language* 16: 311–334.
Wilson, Mark. 2008. *Wandering Significance: An Essay on Conceptual Behavior*. Oxford: Oxford University Press.